Prilagojeno zmanjševanje kompleksnih realnih omrežij

Neli Blagus

doktorska disertacija
predana
Fakulteti za računalništvo in informatiko
kot del izpolnjevanja pogojev za pridobitev naziva
doktor znanosti
s področja
računalništva in informatike

Ljubljana, 2016

PREDHODNA OBJAVA

Izjavljam, da so bili rezultati obravnavane raziskave predhodno objavljeni/sprejeti za objavo v recenzirani reviji ali javno predstavljeni v naslednjih primerih:

[1] N. Blagus, L. Šubelj in M. Bajec. Self-similar scaling of density in complex real-world networks. *Physica A*, 391(8):2794–2802, 2012. doi: 10.1016/j.physa.2011.12.055

[2] N. Blagus, L. Šubelj in M. Bajec. Assessing the effectiveness of real-world network simplification. *Physica A*, 413:134–146, 2014. doi: 10.1016/j.physa.2014.06.065

[3] N. Blagus, L. Šubelj, G. Weiss in M. Bajec. Sampling promotes community structure in social and information networks. *Physica A*, 432:206–215, 2015. doi: 10.1016/j.physa.2015.03.048

Potrjujem, da sem pridobila pisna dovoljenja vseh lastnikov avtorskih pravic, ki mi dovoljujejo vključitev zgoraj navedenega materiala v pričujočo disertacijo. Potrjujem, da zgoraj navedeni material opisuje rezultate raziskav, izvedenih v času mojega podiplomskega študija na Univerzi v Ljubljani.

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*
Neli Blagus
Prilagojeno zmanjševanje kompleksnih realnih omrežij

POVZETEK

Omrežja so pomembno orodje za analizo, razumevanje in prikaz kompleksnih sistemov, kot so na primer družbeno omrežje uporabnikov Facebooka, tehnološko omrežje železniških povezav, biološko omrežje interakcij med beljakovinami in informacijsko omrežje povezav med spletnimi stranmi. Zaradi hitrega razvoja svetovnega spleta in možnosti shranjevanja velikih količin podatkov v zadnjih letih z omrežji opisujemo vse večje sisteme. Velikost omrežij otežuje njihovo razumevanje in prikaz, prav tako je časovno in prostorsko zahtevnejša njihova analiza. Problem za učinkovito analizo predstavljajo tudi hitro spreminjajoča se omrežja ter omrežja z nepopolnimi ali skritimi podatki. Realna omrežja so tako pogosto zmanjšane različice dejanskih sistemov, ki so z omrežji opisani. Z namenom reševanja problema velikosti omrežij in razumevanja razlik med dejanskimi in nepopolnimi sistemi so bili predlagani različni pristopi za zmanjševanje omrežij. Pri procesu zmanjševanja z združevanjem, vzorčenjem ali s preiskovanjem vozlišč in povezav veliko omrežje preoblikujemo v manjše. Hkrati želimo, da se lastnosti osnovnega omrežja z zmanjševanjem čimbolj ohranijo. Zmanjšano omrežje je tako uporabno na primer za hitrejšo analizo, učinkovitejši prikaz ali simulacijo dinamičnih procesov.

V literaturi je analiziranih veliko pristopov za zmanjševanje, redke študije pa medsebojno primerjajo različne pristope. V disertaciji se ukvarjamo z analizo spreminjanja omrežij med zmanjševanjem in primerjavo pristopov za zmanjševanje. Predlagamo mero za oceno učinkovitosti zmanjševanja, ki temelji na uspešnosti ohranjanja lastnosti osnovnega omrežja. S predlagano mero primerjamo več pristopov za zmanjševanje omrežji različnih tipov in velikosti ter opazujemo uspešnost ohranjanja lastnosti pri zmanjšanih omrežjih različnih velikosti. Podrobneje analiziramo spreminjanje gostote omrežij in povezovanje skupin vozlišč med zmanjševanjem. Na podlagi rezultatov oblikujemo shemo za pomoč pri izbiri pristopa za zmanjševanje izbranega omrežja.

Ključne besede: analiza omrežij, realna omrežja, zmanjševanje omrežij, združevanje, vzorčenje, preiskovanje, uspešnost zmanjševanja, skupine vozlišč, gostota omrežja

University *of Ljubljana*
Faculty *of Computer and Information Science*
Neli Blagus
Adapted simplification of complex real-world networks

ABSTRACT

Networks are an important tool for analyzing and visualizing different complex systems. Examples of real-world networks include social network of friends on Facebook, technological network of railways, biological network of interactions between proteins and information networks of hyperlinks between the Web pages. The evolution of the Web and the capability of storing large amounts of data have caused the size of networked systems and their complexity to increase. However, the algorithms for network analysis and visualization appear impractical for addressing very large systems. Furthermore, data about networks are not always complete, their structure may be hidden, or they may change quickly over time. Any network studied in the literature is thus inevitably just a simplified representative of its real-world analogue. For these reasons, understanding how an incomplete system differs from a complete one is crucial. Recently, a number of techniques have been proposed for simplifying complex networks. The simplification is a process, which reduce the size of a large network with merging, sampling or exploration of nodes or links in a network. Simplification techniques are applied to large networks to allow for their faster and more efficient analysis. Since the findings of the analyses and simulations of simplified networks are implied for the original ones, it is of key importance to understand the structural differences between the original networks and their simplified variants.

Network simplification has been extensively investigated from different perspectives. A large number of studies focus on the changes in network properties introduced by simplification. On the other hand, only a few studies compare simplification techniques. In this doctoral thesis, we study the changes of real-world networks introduced by simplification and analyze the differences among simplification techniques. We propose an approach for assessing the effectiveness of simplification. Based on the similarity between original and simplified networks, we compare different simplification tech-

niques. We simplify a number of real-world networks of various types and sizes and explore the preservation of network properties on simplified networks of different sizes. We analyze the changes of network density under the simplification and compare characteristic groups of nodes in original and simplified networks. Based on the findings of the analyses we introduce the scheme for choosing the appropriate simplification technique for a particular network.

Key words: network analysis, real networks, network simplification, merging, sampling, exploration, simplification effectiveness, node groups, network density

KAZALO

I

Uvod

Na družbenem omrežju Facebook uporabniki vsako minuto pošljejo 150.000 sporočil in 100.000 prošenj za prijateljstvo, naložijo 243.000 fotografij in namestijo 14.000 aplikacij. Samo v enem dnevu to pomeni ogromno podatkov o uporabnikih in njihovih navadah. Zbrane podatke raziskovalci uporabljajo na primer pri iskanju vzorcev povezovanj in napovedovanju prihodnjih interakcij med uporabniki ali pri prilagajanju ponudbe reklamnih obvestil. Kako pa te podatke analizirati in prikazati v doglednem času in na razumljiv način?

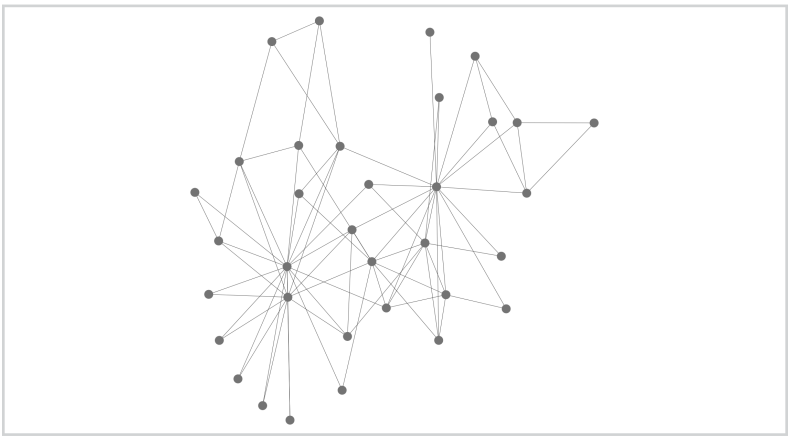
Analiza omrežij (angl. network analysis) [1, 2] izhaja iz teorije grafov [3, 4] in se ukvarja z analizo podatkov, predstavljenih v obliki omrežja. Abstrakcija z omrežji omogoča enostavno predstavitev velikih in kompleksnih sistemov. Orodja analize omrežij pomagajo pri reševanju različnih problemov, od napovedovanja prijateljstev na družbenem omrežju [5] do odkrivanja goljufij v zavarovalništvu [6]. Uporaba analize omrežij tako ni omejena le na računalništvo, temveč sega na mnoga druga področja, vse od fizike, elektrotehnike in biologije do ekonomije in sociologije. Z omrežji lahko na primer opišemo interakcije med geni in beljakovinami (slika 1.2(a)), ponazorimo omrežje železniških povezav (slika 1.2(b)) in računalnike povezane v svetovni splet (slika 1.2(c)) ter analiziramo sodelovanja med znanstveniki (slika 1.2(d)).

1.1 Pregled področja

V teoriji grafov podatke predstavimo kot množico med seboj povezanih objektov. Graf sestavljajo vozlišča, ki predstavljajo objekte, ter povezave med njimi, ki pomenijo različne vrste interakcij med objekti. Število povezav, ki so vezane na vozlišče, označimo s stopnjo vozlišča (angl. degree). Grafu pravimo usmerjen (angl. directed), če je smer povezav med vozlišči pomembna, v nasprotnem primeru je graf neusmerjen (angl. undirected). V usmerjenem grafu imajo vozlišča vhodno (angl. in-degree) in izhodno (angl. out-degree) stopnjo, ki pomenita število povezav, ki kažejo v oziroma iz vozlišča.

Z uporabo grafov so raziskovalci že v 18. stoletju računali napetost v električnih vezjih [7], reševali problem Königsberških mostov [8] in barvanja zemljevida [9]. V 70. letih prejšnjega stoletja so začeli sociologi s pomočjo grafov analizirati odnose med ljudmi [10] (slika 1.1). Vozliščem, ki označujejo ljudi, so pripisali dodatne oznake, na primer ime in starost. Povezavam so dodali vrsto relacije med osebama, na primer prijateljstvo ali sorodstvo. Grafom z dodatnim znanjem o vozliščih in povezavah pravimo omrežja (angl. networks).

Glede na nastanek omrežja razdelimo na realna (angl. real) in naključna



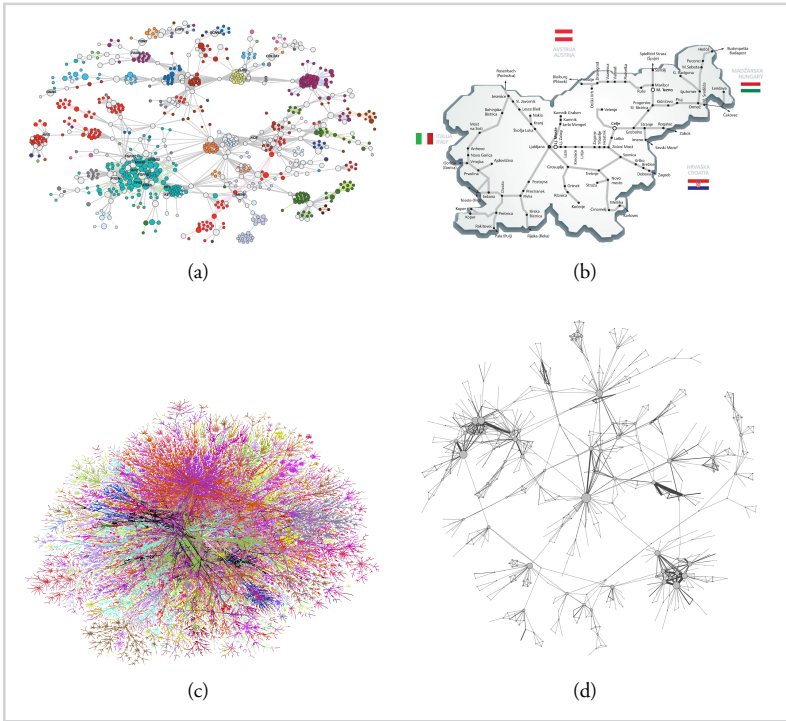
Slika 1.1

Eno najbolj prepoznavnih omrežij prijateljstev je sestavil sociolog Zachary, ko je preučeval odnose med člani karate kluba [10]. Vozlišča omrežja predstavljajo člane kluba, povezave pa ponazarjajo interakcije med njimi.

(angl. random). Razvoj analize omrežij se je začel s proučevanjem realnih omrežij, ki so dobljena iz realnih sistemov. Po drugi strani naključna omrežja [11, 12] dobimo z naključnim dodajanjem povezav med vozlišča. Naključna omrežja se po lastnostih razlikujejo od realnih [13, 14], uporabljamo pa jih predvsem za modeliranje realnih omrežij [15, 16]. V disertaciji se ukvarjamo pretežno z realnimi omrežji, zato se v nadaljevanju osredotočimo nanje.

Poznamo več vrst realnih omrežij, ki jih v grobem razdelimo v štiri skupine: družbena, informacijska, tehnološka in biološka omrežja [17]. V družbenih omrežjih (angl. social networks) vozlišča predstavljajo ljudi, povezave pa pomenijo interakcije med njimi, na primer sorodstvo, poslovni stiki ali prijateljstvo (primer na sliki 1.1). V informacijskih omrežjih (angl. information networks) povezave ustrezajo toku informacij med vozlišči. Primeri informacijskih omrežij so omrežja citiranost med znanstvenimi prispevki, komunikacijska omrežja mobilnih naprav in omrežje povezav med računalniki v svetovnem spletu (primer na sliki 1.2(c)). Tehnološka omrežja (angl. technological networks) predstavljajo infrastrukture, ki so podvržene umetnim ali naravnim vplivom, kot na primer električno, rečno in cestno omrežje (primer na sliki 1.2(b)). Z biološkimi omrežji (angl. biological networks) pa opišemo biološke sisteme, kot so na primer prehranjevalne verige, nevronska ter beljakovinska omrežja (primer na sliki 1.2(a)).

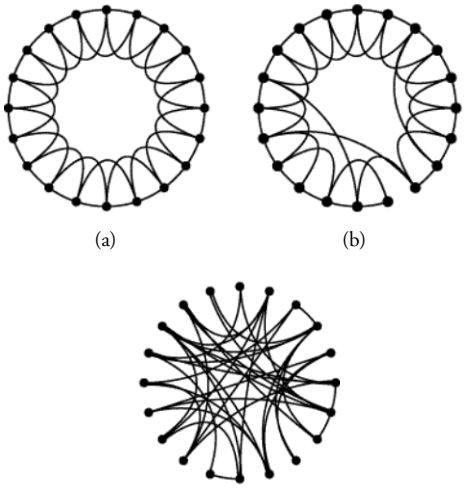
Kljub raznolikosti sistemov, na podlagi katerih so zgrajena realna omrež-



Slika 1.2

Primeri omrežij iz realnega sveta: (a) biološko omrežje interakcij med geni [18], (b) omrežje železniških povezav v Sloveniji [19], (c) internetno omrežje naprav, povezanih v svetovnem spletu [20] ter (d) omrežje sodelovanj med slovenskimi raziskovalci na področju informatike [21].

ja, imajo le-ta veliko skupnih lastnosti. V večini realnih omrežjih sta poljubni vozlišči po najkrajši poti povezani z majhnim številom povezav, prav tako so v lokalni okolici posameznih vozlišč povezave gostejše (slika 1.3). Takšnim omrežjem pravimo omrežja majhnega sveta (angl. small-world networks) [22]. Prvi je pojav majhnega sveta zaznal Milgram pri eksperimentu s pošiljanjem pisem, ki ga je izvedel v 60. letih prejšnjega stoletja. Kmetom v New Yorku je razdelil pisma, namenjena borznemu posredniku v Bostonu. Kmetje so pisma poslali svojim znancem čim bližje Bostonu, ti so pisma spet poslali naprej svojim znancem itd., dokler pisma niso prišla na cilj. Med pošiljanjem se je veliko pisem izgubilo, tista, ki pa so na cilj prišla, so bila v povprečju poslana 6,2-krat. Od tod izhaja teorija šestih stopenj ločenosti (angl. six degrees of separation), ki pravi, da sta katerikoli dve osebi na svetu povezani preko



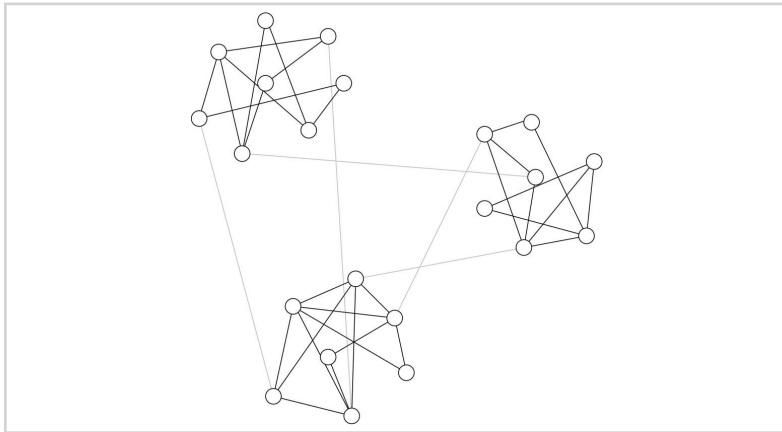
Slika 1.3

Prikaz lastnosti omrežja majhnega sveta [22]: (a) v regularnem omrežju imajo imajo vozlišča enako stopnjo, razdalje med njimi so daljše, lokalno so vozlišča gosto povezana, (b) za omrežja majhnega sveta je značilna majhna povprečna razdalja med vozlišči, lokalno so vozlišča gosto povezana, (c) v naključnem omrežju je vsako vozlišče povezano z naključnim številom drugih vozlišč, razdalje med njimi so kratke, lokalno pa so vozlišča redko povezana.

šest ali manj znancev. Kasneje so raziskave pokazale, da je pojav majhnega sveta prisoten v večini omrežij različnih vrst [22]; na primer družbeno omrežje Facebook ima štiri prostostne stopnje, uporabniki so v omrežju med seboj povezani v povprečju preko manj kot 5 znancev [23].

Večini realnih omrežij je skupna tudi brezlestvičnost (angl. scale-free), kar pomeni, da so stopnje vozlišč v omrežju porazdeljene po potenčnem zakonu (angl. power-law). V brezlestvičnem omrežju obstajajo vozlišča s stopnjo, ki je veliko večja od povprečne stopnje vozlišč omrežja. Brezlestvičnost je v močni korelaciji z odpornostjo (angl. robustness) omrežja, ki pomeni sposobnost delovanja omrežja kljub odstranitvi vozlišč ali povezav. Brezlestvična omrežja vsebujejo vozlišča z visoko stopnjo, imenujemo jih zvezdišča (angl. hubs), ki so povezana z vozlišči z nizko stopnjo. Tako strukturirana omrežja so odporna na odstranitev naključnih vozlišč ali povezav, saj je verjetnost odstranitve vozlišča z zelo visoko stopnjo izredno majhna [22]. Po drugi strani pa namerna odstranitev vozlišča z visoko stopnjo lahko povzroči razpad omrežja na več nepovezanih komponent.

V splošnem za realna omrežja velja, da so zelo redka (angl. sparse) [24], saj

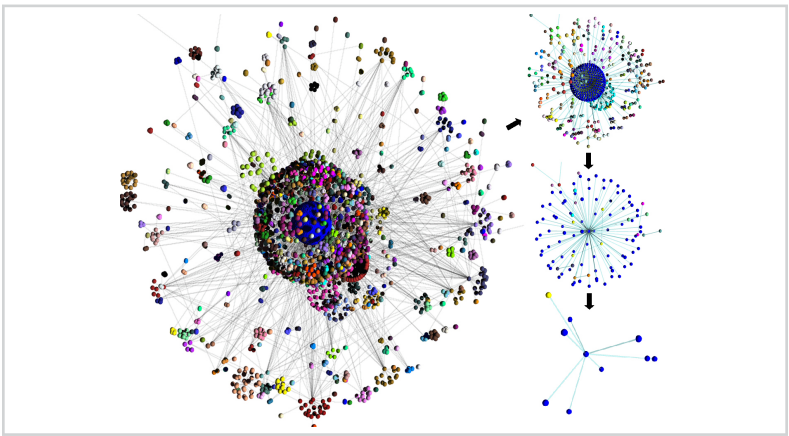


Slika 1.4

Primer omrežja s tremi skupnostmi [27], ki vsebujejo gosto povezana vozlišča, skupnosti med seboj pa so povezane redko.

vsebujejo veliko manj povezav, kot je vseh možnih povezav med vozlišči. Nasprotno velja za gostoto povezav v okolici posameznih vozlišč. Nakopičenost (angl. clustering) [22] meri gostoto omrežja v okolici določenega vozlišča; v družbenem omrežju nakopičenost pomeni verjetnost, da sta prijatelja skupnega prijatelja tudi prijatelja. Tranzitivnost (angl. transitivity) definiramo kot razmerje med številom vseh povezanih trojk vozlišč v omrežju in številom vseh potencialno tranzitivnih trojk. S tranzitivnostjo merimo nakopičenost celotnega omrežja. Omrežja, v katerih je prisoten pojav majhnega sveta, imajo visoko tranzitivnost, saj vsebujejo skupine gosto povezanih vozlišč. V brezlestvičnih omrežjih je nakopičenost porazdeljena po potenčnem zakonu. Vozlišča z majhno stopnjo se združujejo v goste skupine, ki so med seboj povezane preko zvezdišč. V družbenem omrežju bi to pomenilo, da se ljudje v majhnih skupinah poznajo med seboj, več manjših skupin pa je med seboj povezanih prek manjšega števila znancev. Gosto povezanim manjšim skupinam vozlišč, ki so med seboj povezane redko, pravimo skupnosti (angl. community) (slika 1.4). Skupnosti ustrezajo na primer skupinam oseb s podobnimi interesi v družbenem omrežju [25], znanstvenim področjem v omrežju citiranj [26] in funkcionalnim skupinam v metaboličnih omrežjih [26].

Odkrivanje skupnosti (angl. community detection) [27] je v zadnjih letih zelo raziskano področje analize omrežij [29, 30]. Poznavanje strukture skupnosti omogoča lažje razumevanje obnašanja in spreminjanja sistema, ki je



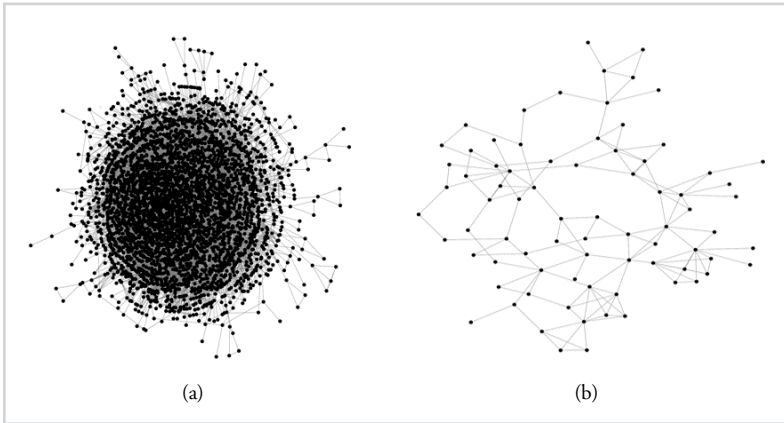
Slika 1.5

Primer renormalizacije internetnega omrežja [28], kjer so v nadvozlišča združena vozlišča na razdalji manj kot 3.

z omrežjem opisan, saj imajo skupnosti močan vpliv na primer na dinamične procese v omrežju [31]. Poleg skupnosti se vozlišča v omrežju povezujejo tudi v druge vrste skupin [32, 33]. Šubelj in Bajec [34] sta pokazala, da veliko realnih omrežij vsebuje module (angl. modules), sestavljene iz vozlišč, ki med seboj niso nujno povezana, so pa podobno povezana z ostalimi vozlišči v omrežju. Moduli ustrezajo na primer dokumentom s podobno vsebino v spletnih omrežjih in besedam iste vrste v leksikalnih omrežjih.

Pri analizi nas pogosto zanimajo najpomembnejša vozlišča v omrežju, ki jih določimo z merami središčnosti (angl. centrality measures). Z vmesno središčnostjo (angl. betwenness centrality) [35] merimo, kolikokrat posamezno vozlišče služi kot most na najkrajši poti med vsemi možnimi pari vozlišč v omrežju. Preprostejša mera pomembnosti vozlišč je stopnja, ki pomeni število povezav, vezanih na vozlišče. V nekaterih omrežjih so izraziti vzorci mešanja (angl. mixing patterns) [36] med podobnimi oziroma različnimi vozlišči; na primer mešanje stopenj (angl. degree mixing) meri korelacijo med stopnjami vozlišč, ki so povezana. Mešanje je pozitivno kolerirano (angl. assortative), če se vozlišča z visoko stopnjo povezujejo med seboj, kar velja na primer za družbena omrežja. V bioloških in tehnoloških omrežjih se vozlišča z visoko stopnjo povezujejo z vozlišči z manjšo stopnjo, kar imenujemo negativno kolerirano (ang. disassortative) mešanje stopenj.

Pomembna lastnosti, ki nam pomaga razumeti razvoj in spreminjanje



Slika 1.6

Primer zmanjševanja omrežja: (a) osnovno omrežje in (b) zmanjšano omrežje.

omrežij, je samopodobnost (angl. self-similarity). V samopodobnih omrežjih obstaja potenčno razmerje med velikostjo omrežja in številom njegovih delov, ki jih dobimo med procesom renormalizacije (angl. renormalization). Pri iterativnem procesu renormalizacije združujemo vozlišča v nadvozlišča (angl. supernodes), ki jih povežemo z nadpovezavami (angl. superlinks), če vsebujejo med seboj povezana vozlišča (slika 1.5). Vozlišča združujemo glede na različne lastnosti, na primer glede na razdaljo med vozlišči [37] ali pripadnost isti skupnosti [38].

Renormalizacija predstavlja primer zmanjševanja omrežja (angl. network simplification). V splošnem je zmanjševanje postopek, pri katerem iz velikega omrežja oblikujemo manjše omrežje (slika 1.6), primernejše za hitrejšo analizo, učinkovitejši prikaz in lažje razumevanje procesov, ki potekajo v omrežju. Običajno želimo omrežje zmanjšati in hkrati zagotoviti podobnost med osnovnim in zmanjšanim omrežjem. Slednje nam omogoča uporabo rezultatov analize zmanjšanega omrežja na velikem omrežju.

Raziskovalci so predlagali veliko pristopov za zmanjševanje, ki jih v grobem razdelimo v tri skupine: zmanjševanje z združevanjem (angl. merging), zmanjševanje z vzorčenjem (angl. sampling) in zmanjševanje s preiskovanjem (angl. exploration). Pri zmanjševanju z združevanjem zmanjšano omrežje dobimo z združevanjem vozlišč [37] ali povezav [39], primer predstavlja zgoraj opisani proces renormalizacije. Pri zmanjševanju z vzorčenjem je zmanjšano omrežje sestavljeno iz vozlišč, ki jih iz osnovnega omrežja izberemo naključ-

no ali sorazmerno glede na izbrano lastnost [40]; na primer pri naključnem izbiranju vozlišč (angl. random node selection) v vzorec naključno izberemo vozlišča, pri naključnem izbiranju glede na stopnjo (angl. random node selection based on degree) pa z večjo verjetnostjo izberemo vozlišča z večjo stopnjo. Podobno lahko v vzorec izbiramo povezave [40]. V zadnjo skupino zmanjševanj s preiskovanjem spadajo pristopi, kjer zmanjšano omrežje predstavlja podomrežje celotnega omrežja [40]. Zmanjšano omrežje dobimo tako, da z začetkom v naključno izbranem vozlišču preiskujemo njegovo lokalno okolico in v zmanjšano omrežje vključimo določeno število preiskanih vozlišč. Primera zmanjševanja s preiskovanjem sta preiskovanje v širino [41] in preiskovanje z naključnim sprehodom [42]. Lahko pa v zmanjšanem omrežju ohranimo vsa vozlišča in samo najpomembnejše povezave med njimi [43], ki jih izberemo glede na frekvenco pojavljanja v najkrajših poteh med vozlišči [44] ali v minimalnih vpetih poddrevesih omrežja [45].

1.2 Motivacija in cilji

V začetkih teorije omrežij so se raziskovalci ukvarjali z majhnimi sistemi, sestavljenimi iz nekaj deset ali sto objektov. Takšna omrežja je preprosto analizirati z obstoječimi algoritmi teorije grafov, jih narisati na papir ali prikazati na zaslonu. V zadnjih letih, predvsem z razvojem svetovnega spleta in možnostjo shranjevanja velikih količin podatkov, z omrežji opisani sistemi postajajo vse večji. Algoritmi za analizo so časovno in prostorsko preveč zahtevni, omrežja z nekaj milijoni vozlišč pa je praktično nemogoče prikazati na zaslonu. Poleg velikih omrežij predstavljajo problem za učinkovito analizo tudi hitro spreminjajoča se omrežja ter omrežja z nepopolnimi ali skritimi podatki.

Zmanjševanje omrežij se je sprva uporabljalo za obvladovanje velikosti omrežij predvsem pri shranjevanju, poudarek raziskav je bil na stiskanju omrežij (angl. compression) za lažje shranjevanje [46, 47]. Z rastjo omrežij so se razvijali tudi pristopi za zmanjševanje kot pomoč pri hitrejši analizi [48, 49] in učinkovitejšemu prikazu omrežij [50, 51]. Hkrati pa so pristopi za zmanjševanje omogočili tudi primerjavo velikih in zmanjšanih omrežij. Z analiziranjem razlik in podobnosti med njimi tako razumemo, kako se razlikujeta veliko omrežje z nepopolnimi podatki in omrežje, ki je na voljo za analizo. Slednje na podoben način pomaga tudi pri analizi in razumevanju hitro spreminjajočih se omrežij ali omrežij, ki vsebujejo skrite in manjkajoče podatke. V zadnjem času se večina študij na temo zmanjševanja osredotoča na zmanjševanje dolo-

čenega tipa omrežij [45, 52, 53] ali analizo spreminjanja posameznih lastnosti med zmanjševanjem [54–56]. Redke študije pa medsebojno primerjajo delovanje različnih pristopov za zmanjševanje. Poznavanje razlik med pristopi omogoča smotrno izbiro pristopa za zmanjševanje, prilagojeno namenom in ciljem nadaljnje analize omrežja.

V disertaciji se ukvarjamo z analizo spreminjanja omrežij med zmanjševanjem. Študije so pokazale, da se omrežja z zmanjševanjem spreminjajo, od uporabljenega pristopa pa je odvisno, kako se spremenijo posamezne lastnosti [40, 57]. Pristopi za zmanjševanje se namreč razlikujejo po primernosti za ohranjanje določenih lastnosti [54, 58]. Pri našem delu predpostavimo, da tudi tip in velikost osnovnega omrežja vplivata na to, kako dobro se posamezne lastnosti z zmanjševanjem ohranijo. Domnevamo, da so si omrežja določenega tipa dovolj podobna, da lahko uspešno ohranimo njihove lastnosti z enakimi pristopi. Poleg tega predpostavimo, da na uspešnost ohranjanja lastnosti vpliva tudi velikost zmanjšanega omrežja. Večje kot je zmanjšano omrežje, bolj povzame lastnosti osnovnega. Nasprotno velja za manjše zmanjšano omrežje. V disertaciji predlagamo mero za oceno uspešnosti zmanjševanja, s katero preverimo zgornje predpostavke. S pomočjo predlagane mere analiziramo, kako se z zmanjševanjem spreminjajo lastnosti omrežij različnih tipov in velikosti. Opazujemo tudi, kako velikost zmanjšanih omrežij vpliva na uspešnost zmanjševanja. Podrobneje analiziramo spreminjanje gostote omrežij pri zmanjševanju z združevanjem vozlišč in povezovanje skupin vozlišč pri zmanjševanju družbenih in informacijskih omrežij.

Poleg analize spreminjanja omrežij med zmanjševanjem se v disertaciji ukvarjamo s primerjavo pristopov za zmanjševanjem. V splošnem pristopi za zmanjševanje z vzorčenjem ohranijo lastnosti osnovnih omrežij slabše kot pristopi za zmanjševanje s preiskovanjem [40]. Zmanjševanja z vzorčenjem namreč ustvarijo manj povezana omrežja, sestavljena iz več nepovezanih komponent. Po drugi strani pa pristopi za zmanjševanje z združevanjem ustvarijo povezana omrežja, ki bi lahko bila bolj podobna osnovnim. Predvidevamo, da lastnosti osnovnih omrežij ohranijo bolje. Slednjo hipotezo preverimo s primerjavo pristopov iz obeh skupin. Rezultati nedavne študije pokažejo, da z vključevanjem dodatnih povezav v zmanjšano omrežje znatno izboljšamo delovanje naključnega izbiranja povezav [59]. Izboljšan pristop bolje kot drugi ohrani nekatere lastnosti osnovnih omrežij. Predpostavimo, da lahko na podoben način izboljšamo tudi delovanje pristopov za zmanjševanje s preiskovanjem. S pomočjo predlagane mere primerjamo različne pristope za zmanjše-

vanje z vzorčenjem in s preiskovanjem ter opazujemo ohranjanje lastnosti pri različnih velikostih zmanjšanih omrežij.

Ob veliki množici različnih pristopov za zmanjševanje se poraja vprašanje, kako izbrati primeren pristop za zmanjševanje določenega omrežja. Predpostavimo, da izbira temelji predvsem na tem, katere lastnosti želimo z zmanjševanjem ohraniti. Na podlagi rezultatov vseh primerjav tako oblikujemo shemo za pomoč pri izbiri pristopa za zmanjševanje.

1.3 Prispevki k znanosti

V disertaciji predstavimo tri glavne prispevke k znanosti:

- *Mera za oceno uspešnosti zmanjševanja:* predlagamo mero, s katero primerjamo pristope za zmanjševanje na podlagi podobnosti med osnovnim in zmanjšanim omrežjem. Analiziramo primernost pristopov za ohranjanje posameznih lastnosti omrežja. Opazujemo, kako na uspešnost zmanjševanja vplivajo tip in velikost osnovnih ter velikost zmanjšanih omrežij.

- *Analiza spreminjanja omrežij med zmanjševanjem:* opazujemo, kako se z zmanjševanjem spreminja gostota omrežij različnih tipov in velikosti. Analiziramo, kako se spreminja povezovanje skupin vozlišč pri zmanjševanju družbenih in informacijskih omrežij.

- *Shema za izbiro pristopa za zmanjševanje:* rezultate analiz združimo v shemo za pomoč pri izbiri pristopa za zmanjševanje. S pomočjo sheme izberemo pristop glede na to, katere lastnosti želimo z zmanjševanjem ohraniti.

1.4 Pregled vsebine

Disertacija temelji na treh objavljenih člankih, katerih glavno idejo in prispevke predstavimo v poglavju 2. Članki so v disertacijo vloženi v poglavjih 3, 4 in 5. V prvem predlagamo mero za oceno uspešnosti zmanjševanja in med seboj primerjamo pristope za zmanjševanje z združevanjem in vzorčenjem. V drugem članku opazujemo spreminjanje gostote ter v tretjem spreminjanje skupin vozlišč med zmanjševanjem. V poglavju 6 analiziramo izboljšavo pristopov za zmanjševanje s preiskovanjem. Med seboj primerjamo pristope za

zmanjševanje z vzorčenjem in s preiskovanjem ter oblikujemo shemo za pomoč pri izbiri pristopa za zmanjševanje izbranega omrežja. Disertacijo zaključimo s povzetkom rezultatov in predlogi za nadaljnje delo v poglavju 7.

2

Pregled objavljenih del

V vseh treh člankih, ki sestavljajo jedro tega dela, avtorica disertacije nastopam kot prvi avtor. Moje delo pri člankih je obsegalo zasnovo in izvedbo eksperimentov, analizo in interpretacijo rezultatov ter pisanje člankov.

V prvem članku [60] predlagamo mero, s katero primerjamo pristope za zmanjševanje na podlagi podobnosti med osnovnim in zmanjšanim omrežjem. Analiziramo vpliv tipa in velikosti omrežij na uspešnost zmanjševanja, kjer boljša uspešnost pomeni večjo podobnost med osnovnim in zmanjšanim omrežjem. Omrežja zmanjšamo na velikosti med 1 % in 50 % osnovnih omrežij. Izkaže se, da velikost zmanjšanega omrežja vpliva na uspešnost zmanjševanja, tip in velikost osnovnega omrežja pa ne. Večja zmanjšana omrežja bolje povzamejo lastnosti osnovnih omrežij kot manjša. Ker pa je običajno cilj zmanjševanja omrežje čimbolj zmanjšati, pokažemo, da pri zmanjšanih omrežjih velikosti okrog 10 % osnovnih omrežij dosežemo kompromis med velikostjo omrežja in ohranjanjem lastnosti osnovnih omrežij. V drugem delu študije se osredotočimo na zmanjšana omrežja velikosti 10 % ter opazujemo ohranjanje posameznih lastnosti pri zmanjševanju z različnimi pristopi. Analiza pokaže, da so pristopi za zmanjševanje z vzorčenjem uspešnejši od pristopov, ki vozlišča združujejo.

V naslednjih dveh člankih analiziramo spreminjanje gostote in povezovanje skupin vozlišč med zmanjševanjem. V drugem članku [38] pokažemo, da obstaja potenčno razmerje med velikostjo omrežij in njihovo gostoto. Razmerje velja ne glede na tip omrežij in izbran pristop zmanjševanja z združevanjem. Rezultati nadgradijo študijo, ki analizira potenčno razmerje med velikostjo in gostoto omrežij različnih tipov in velikosti [24]. V tretjem članku [61] pokažemo, da se povezovanje skupin vozlišč med zmanjševanjem družbenih in informacijskih omrežij spremeni. Ne glede na tip omrežja in pri večini pristopov za zmanjševanje postanejo skupine z zmanjševanjem gostejše.

V poglavju 6 se osredotočimo na primerjavo pristopov za zmanjševanje z vzorčenjem in s preiskovanjem. Analize so pokazale, da se z indukcijo, kjer dodamo povezave osnovnega omrežja v zmanjšano omrežje, izboljša delovanje naključnega izbiranja povezave [59]. Po zgledu slednje študije korak indukcije vpeljemo v pristopa zmanjševanja s preiskovanjem. Med seboj primerjamo pristope na zmanjšanih omrežjih različnih velikosti. Ocenimo jih z mero, predlagano v poglavju 3. Analiza odkrije več razlik med pristopi zmanjševanja z indukcijo in brez nje. Pristopi z indukcijo izboljšajo delovanje pri ohranjanju porazdelitve stopenj vozlišč in nakopičenosti, zmanjšana omrežja imajo višjo povprečno stopnjo ter so gostejša. Podrobneje analiziramo tudi uspešnost

pristopov na zmanjšanih omrežjih velikosti 10 % osnovnih omrežij, rezultate vseh analiz pa uporabimo pri oblikovanju sheme za pomoč pri izbiri pristopa za zmanjševanje velikega omrežja.

2.1 Primerjava pristopov za zmanjševanje

Kljub velikemu številu raziskav na temo zmanjševanja omrežij je malo takih, ki med seboj primerjajo različne pristope za zmanjševanje. Leskovec in sodelavci [40] so opazovali spreminjanje lastnosti omrežij med zmanjševanjem ter merili podobnost med osnovnimi in zmanjšanimi omrežji s pomočjo D-statistike Kolmogorov-Smirnova. Lee in sodelavci [57] so iskali vzorce, po katerih se spreminjajo določene lastnosti med zmanjševanjem. Hübler in sodelavci [62] so primerjali pristope za zmanjševanje na podlagi povprečne razlike med lastnostmi osnovnih in zmanjšanih omrežij, podobno sta Doer in Blenn [63] primerjala dejanske vrednosti lastnosti osnovnih in zmanjšanih omrežij. Toivonen in sodelavci [64] so predlagali več pristopov za kompresijo uteženih omrežij in jih med seboj primerjali glede na čas izvajanja in zahtevnost shranjevanja. Vse naštete študije se osredotočajo predvsem na primerjavo na novo predlaganega pristopa z obstoječimi [62] ter analizirajo delovanje pristopov na manjši množici omrežij [40, 57, 63].

V članku, vloženem v poglavju 3, analiziramo in primerjamo pristope za zmanjševanje. Opazujemo, kako na uspešnost zmanjševanja vpliva tip in velikost osnovnih ter velikosti zmanjšanih omrežij. V analizo zajamemo 30 omrežij različnih tipov in velikosti ter analiziramo, kako se njihove lastnosti spreminjajo med zmanjševanjem. Pri analizi se osredotočimo na tri vprašanja: kako meriti podobnost med osnovnim in zmanjšanim omrežjem, kako izbrati velikost zmanjšanega omrežja ter kako primerjati pristope med seboj.

Mera za oceno uspešnosti zmanjševanja

Pri zmanjševanju z vzorčenjem in s preiskovanjem je zmanjšano omrežje sestavljeno iz vozlišč in povezav osnovnega omrežja. Obe skupini pristopov omogočata nastavljanje velikosti zmanjšanega omrežja; zmanjšano omrežje lahko vsebuje želeno število vozlišč ali določen odstotek vseh vozlišč osnovnega omrežja. Pri zmanjševanju z združevanjem velikosti zmanjšanega omrežja ne moremo nastaviti vnaprej. Lahko pa pri združevanju glede na razdaljo spreminjamo razdaljo; pri manjši razdalji nadvozlišče vsebuje manj vozlišč in je zato zmanjšano omrežje večje, večja razdalja pa pomeni več vozlišč v nadvozlišču, zato je

zmanjšano omrežje manjše.

Podobnost med osnovnim in zmanjšanim omrežjem merimo na podlagi različnih globalnih in lokalnih lastnosti. Analizirane globalne lastnosti so gostota, mešanje stopenj vozlišč [65] ter tranzitivnost [66], lokalne pa porazdelitev stopnje, vhodne ter izhodne stopnje, nakopičenosti [22] in vmesne središčnosti [67]. Globalne lastnosti med osnovnim in zmanjšanim omrežjem primerjamo s Spearmanovim koeficientom korelacije ρ, porazdelitve lokalnih lastnosti pa z D-statistiko Kolmogorov-Smirnova. Nato definiramo mero, s pomočjo katere določimo najbolj primerno velikost zmanjšanega omrežja za ohranjanje posameznih lastnosti. Omrežja zmanjšamo na velikosti med 1 % in 50 % osnovnih omrežij, pri zmanjševanju z združevanjem nastavimo razdalje med vozlišči na celoštevilske vrednosti med 2 in 6. Za vsako omrežje in vsako lastnost razvrstimo velikosti glede na točnost ohranjanja lastnosti. Velikost, kjer je določena lastnost najbolje ohranjena, dobi rang 0, naslednja velikost rang 1 in tako dalje. Range za vsako velikost seštejemo ter vsoto delimo z vrednostjo največje možne vsote. Mera A je definirana kot:

$$A = \frac{1}{(n_s - 1) \cdot n_p} \sum_{i=1}^{n_p} r_i, \tag{2.1}$$

kjer je n_s število različnih velikosti zmanjšanih omrežij, n_p število lastnosti, i indeks lastnosti, kjer vrstni red lastnosti ni pomemben, in r_i rang i–te lastnosti. A je normalizirani skupni rang za neko velikost zmanjšanega omrežja pri določeni lastnosti.

Predlagana mera temelji na rangiranju pristopov, da lahko določimo vrstni red pristopov za primernost ohranjanja posameznih lastnosti. Namesto rangov bi lahko pri ocenjevanju uporabili tudi kar vrednosti sprememb lastnosti med zmanjševanjem ali vrednost statistik, ki jih uporabimo za primerjanje lastnosti. Ker pa za poljubno omrežje ne poznamo pomembnosti posamezne lastnosti in razlike lastnosti v različnih omrežjih nimajo istega pomena, je pri ocenjevanju pristopov na večji množici omrežij in lastnosti primernejše rangiranje.

V prvem delu analize s pomočjo mere pokažemo, da se lastnosti omrežja bolje ohranijo pri večjih zmanjšanih omrežjih. Cilj zmanjševanja je omrežje čimbolj zmanjšati, zato si želimo poiskati kompromis med velikostjo zmanjšanega omrežja in ohranjanjem podobnosti z osnovnim omrežjem. Najboljšo velikost zmanjšanega omrežja definiramo kot lokalni minimum mere A, dosežen pri najmanjši velikosti zmanjšanega omrežja. Rezultati kažejo, da je pri

zmanjševanju z vzorčenjem najboljša velikost zmanjšanih omrežij med 1 % in 15 % osnovnih omrežij, kar se sklada z rezultati podobnih študij [40, 63]. V drugem delu študije podrobneje opazujemo zmanjšana omrežja velikosti 10 % osnovnih omrežij. S predlagano mero, kjer namesto velikosti zmanjšanih omrežij primerjamo pristope, analiziramo primernost pristopov za ohranjanje posameznih lastnosti omrežij. V splošnem sta najbolj učinkovita pristopa naključno izbiranje vozlišč glede na stopnjo ter preiskovanje v širino. Pristopi, ki delujejo po principu združevanja, lastnosti ohranijo najslabše.

Analiza pokaže, da velikost zmanjšanih omrežij vpliva na uspešnost zmanjševanja, medtem ko tip in velikost osnovnih omrežij na uspešnost nimata večjega vpliva. Pri podrobnejšem pregledu rezultatov smo opazili tudi izjeme. Lastnosti omrežij velikosti med 50.000 in 200.000 vozlišč se pri zmanjševanju bolje ohranijo pri manjših zmanjšanih omrežjih. Nasprotno pa velja za večja omrežja velikosti med 200.000 in 500.000 vozlišči. Slednje domneve smo v članku neustrezno statistično preverili. Analiza vpliva tipa in velikosti osnovnega omrežja na uspešnost zmanjševanja tako predstavlja možnost nadaljnjega dela. Za potrditev domnev in ustrezno statistično preverjanje bi bilo potrebno vzorce omrežij pravilno izbrati iz populacije ter uporabiti primernejše statistične teste.

Mero za oceno uspešnosti zmanjševanja ter primerjavo pristopov za zmanjševanje predstavimo v članku [60], v disertacijo vloženem v poglavju 3. Analiza predstavlja pomemben doprinos k razumevanju in smotrni uporabi pristopov za zmanjševanje realnih omrežij.

Ohranjanje globalnih lastnosti med zmanjševanjem

V članku, vstavljenem v poglavje 3 za primerjavo globalnih lastnosti med osnovnim in zmanjšanim omrežjem uporabimo Spearmanov koeficient korelacije. Za vsak pristop zmanjševanja za vsa omrežja skupaj analiziramo, kako so povezane vrednosti gostote, mešanja stopenj in tranzitivnosti v osnovnih in zmanjšanih omrežjih. Poznavanje te povezanosti omogoča sklepanje o vrednosti lastnosti osnovnega omrežja iz vrednosti lastnosti zmanjšanega omrežja. Podobnost omrežij glede na globalne lastnosti v članku ne obravnavamo, zato na tem mestu predstavimo rezultate dodatne analize, kjer merimo podobnost globalnih lastnosti v osnovnih in zmanjšanih omrežij. Primerjavo izvedemo na podlagi dejanskih vrednosti lastnosti. Analiza podobnosti in povezanosti globalnih lastnosti v osnovnih in zmanjšanih omrežjih sta pomembni pri razu-

Tabela 2.1

Kratice pristopov za zmanjševanje.

Kratica	Pristop
RN	Naključno izbiranje vozišč
RD	Naključno izbiranje vozišč glede na stopnjo
RL	Naključno izbiranje povezav
BF	Preiskovanje v širino
CG	Združevanje vozlišč glede na razdaljo
BP	Združevanje vozlišč glede na skupnosti

mevanju spreminjanja omrežij med zmanjševanjem. Pri zmanjševanju želimo, da so lastnosti zmanjšanega omrežja čimbolj podobne lastnostim osnovnega omrežja, če pa tega ne moremo doseči, nas zanima, ali so lastnosti vsaj na kakšen način povezane.

Za določitev najboljše velikosti zmanjšanih omrežij na podlagi globalnih lastnosti za vsak pristop in vsako lastnost razvrstimo velikosti glede na absolutno razliko med vrednostjo lastnosti v osnovnem omrežju in vrednostjo lastnosti v zmanjšanem omrežju. Nato velikostim dodelimo range in jih ocenimo z mero A po formuli 2.1. Rezultati za povprečno oceno A za vse globalne lastnosti pri zmanjševanju z vzorčenjem in s preiskovanjem so prikazani na sliki 2.1(a). Na slikah in v nadaljevanju besedila za posamezne pristope uporabljamo kratice, pojasnjene v tabeli 2.1. Opazimo, da so pri pristopih RD in BF globalne lastnosti najbolje ohranjene pri večjih zmanjšanih omrežjih. Rezultat je primerljiv z rezultati iz slike 3.3(c). Po drugi strani pa pristopa RN in RL dobro ohranita globalne lastnosti že pri manjših zmanjšanih omrežjih pri $s = 0.01$. V analizi članka so najbolje ohranjene pri velikosti $s = 0.25$ (slika 3.3(c)). Rezultati za povprečno oceno A pri zmanjševanju z združevanjem so prikazani na sliki 2.1(b). Globalne lastnosti so najbolje ohranjene pri velikosti $c = 3$, kar je konsistentno z rezultati iz slike 3.4(b).

Podrobneje smo analizirali tudi, katere so najboljše velikosti zmanjšanih omrežij za ohranjanje posameznih globalnih lastnosti. Rezultati so prikazani v tabeli 2.2. Pri zmanjševanju z združevanjem sta tako gostota kot mešanje stopenj najbolje ohranjeni pri $c = 2$, tranzitivnost pa pri manjših omrežjih pri $c = 4$. Pri zmanjševanju z RD in BF so globalne lastnosti najbolje ohranjene pri večjih zmanjšanih omrežjih. Po drugi strani pa se gostota z RN in RL bolje

Tabela 2.2

Najboljše velikosti c pri zmanjševanju z združevanjem in s pri zmanjševanju z vzorčenjem in s preiskovanjem za ohranjanje globalnih lastnosti s pripadajočimi ocenami A.

Lastnost	CG	RN	RD	BF	RL
Gostota	2 (0.31)	0.10 (0.41)	0.50 (0.06)	0.50 (0.03)	0.05 (0.45)
Mešanje stopenj	2 (0.075)	0.01 (0.37)	0.50 (0.20)	0.50 (0.21)	0.50 (0.13)
Tranzitivnost	4 (0.44)	0.50 (0.16)	0.50 (0.16)	0.50 (0.13)	0.01 (0.39)

Tabela 2.3

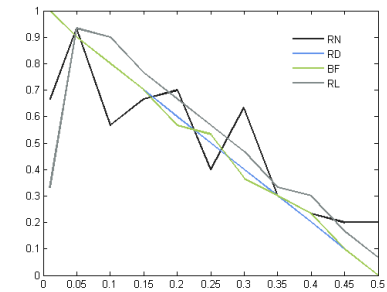
Najboljša dva in najslabši pristop za ohranjanje globalnih lastnosti omrežij.

Lastnost	Najboljši	Drugi najboljši	Najslabši
Gostota	RN	RL	BF
Mešanje stopenj	BF	RL	BP
Tranzitivnost	RN	BF	CG

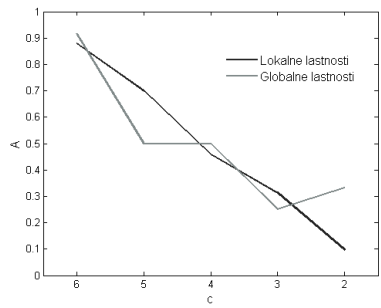
ohrani pri manjših zmanjšanih omrežjih, podobno velja za mešanje stopenj pri zmanjševanju z RN in za tranzitivnost pri zmanjševanju z RL. Do razlik z rezultati iz analize članka v tabeli 3.4 pride pri mešanju stopenj z RN in tranzitivnostjo z RL, ki se v prejšnji analizi ohranita bolje v večjih zmanjšanih omrežjih.

Rezultati kažejo, da so v nekaterih primerih za ohranjanje lastnosti boljša zmanjšana omrežja manjših velikosti. Možna razlaga slednjega bi bila, da imajo omrežja neko lastnost dokaj enakomerno razporejeno povsod po omrežju. Zato lahko tudi relativno majhno podomrežje oziroma kos omrežja dobro povzame lastnost celotnega omrežja.

Na koncu preverimo, kateri so najboljši in najslabši pristopi za ohranjanje posameznih globalnih lastnosti pri zmanjšanih omrežij 10% velikosti osnovnih omrežij (poglavje 3.3.2). Rezultati so prikazani v tabeli 2.3. RN se izkaže kot najboljši pristop za ohranjanje gostote in tranzitivnosti, BF pa za ohranjanje mešanja stopenj. Najslabša za ohranjanje globalnih lastnosti sta pristopa zmanjševanja z združevanjem in BF. Razlike v primerjavi z analizo iz članka v poglavju 3 so predvsem pri drugih najboljših pristopih (tabela 3.5).



(a)



(b)

Slika 2.1

Razdalja med osnovnimi in zmanjšanimi omrežji pri različnih pristopih zmanjševanja merjena s povprečno oceno A preko vseh omrežij (a) za globalne lastnosti pri zmanjševanju z vzorčenjem in s preiskovanjem; (a) za lokalne in globalne lastnosti pri zmanjševanju z združevanjem.

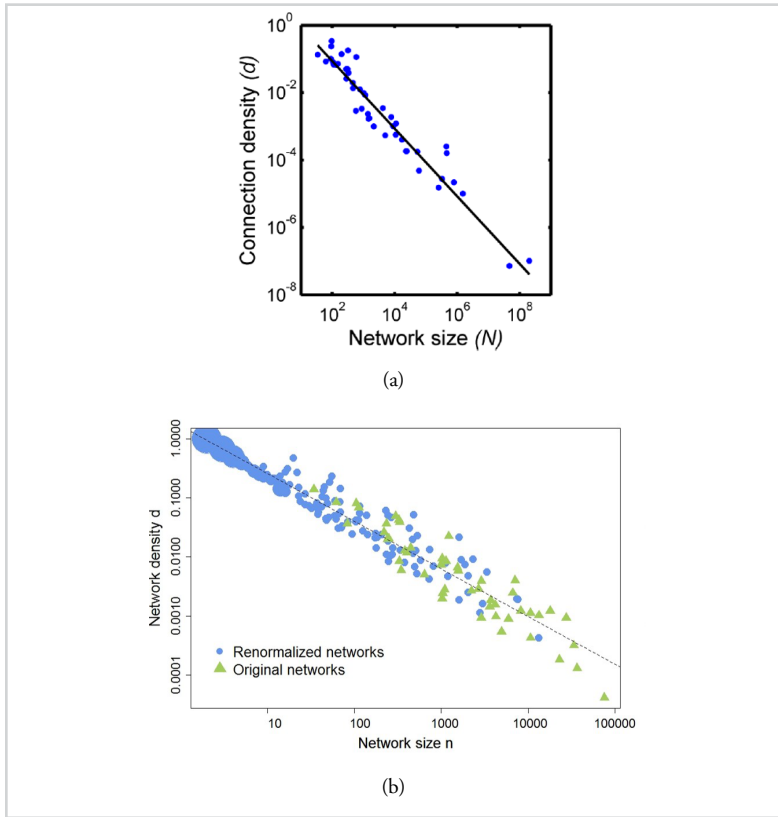
2.2 Spreminjanje omrežij med zmanjševanjem

Študije analizirajo zmanjševanje omrežij iz različnih vidikov. Nekatere se osredotočajo na zmanjševanje določene vrste omrežij, kot na primer zmanjševanje družbenih [45], brezlestvičnih [52] ali usmerjenih omrežij [53]. Druge analizirajo spreminjanje posameznih lastnosti med zmanjševanjem, kot na primer spreminjanje porazdelitve stopenj vozlišč [54], nakopičenosti [58], povezanosti omrežja [55] ali povezovanje skupin [56]. Kljub vsem študijam še zdaleč ne vemo vsega o tem, kaj vpliva na uspešnost zmanjševanja ter kateri pristopi so bolj primerni za ohranjanje posameznih lastnosti. V dveh objavljenih člankih [38, 61], predstavljenih v poglavju 4 in 5, podrobneje raziščemo spreminjanje gostote omrežja in povezovanje skupin vozlišč med zmanjševanjem. Obe analizi predstavljata pomemben doprinos k razumevanju spreminjanja omrežij med zmanjševanjem.

Spreminjanje gostote

Po definiciji je omrežje samopodobno, če obstaja potenčno razmerje med njegovo velikostjo in številom nadvozlišč pri zmanjševanju z združevanjem. Samopodobnost pa so raziskovalci analizirali tudi v kontekstu drugih lastnosti, na primer samopodobna velikost skupnosti [68], porazdelitev stopenj vozlišč [69] in maksimalna stopnja vozlišč [70]. Laurienti in sodelavci [24] so opazovali gostoto različnih omrežij ter odkrili potenčno razmerje med velikostjo in gostoto omrežij različnih tipov in velikosti. V članku, vloženem v poglavju 4, omenjeno študijo nadgradimo z analizo spreminjanja gostote omrežij med zmanjševanjem.

Naj bo $d = \frac{2e}{n(n-1)}$ gostota omrežja, kjer je n število vozlišč in e število povezav v omrežju. Gostoto d lahko zapišemo tudi kot potenčno funkcijo n: $d = c \cdot n^{-\gamma}$, kjer je γ eksponent renormalizacije in c konstanta. S pomočjo determinacijskega koeficienta R^2 merimo ujemanje med gostoto in velikostjo omrežja, s Spearmanovim koeficientom korelacije ρ pa spreminjanje gostote z manjšanjem velikosti omrežja. Rezultati kažejo, da lahko razmerje med gostoto in velikostjo osnovnih omrežij opišemo s potenčnim razmerjem $d \approx n^{-1}$; rezultati so statistično značilni pri vrednosti $p = 0,01$. Pri upoštevanju osnovnih in zmanjšanih omrežij je razmerje v enačbi še močnejše (slika 2.2). Samo osnovna omrežja so precej različna, medtem ko so zmanjšana omrežja ustvarjena iz osnovnih z enakimi pristopi, kar pojasni močnejše potenčno razmerje



Slika 2.2

Potenčno razmerje med velikostjo in gostoto (a) osnovnih omrežij različnih tipov in velikosti [24] ter (a) zmanjšanih omrežij z združevanjem [38].

med gostoto in velikostjo ob upoštevanju tako osnovnih kot tudi zmanjšanih omrežij. Prav tako se izkaže, da razmerje velja za več pristopov zmanjševanja z združevanjem realnih omrežij, ne velja pa za naključna omrežja.

Spreminjanje skupin vozlišč

Številna realna omrežja so sestavljena iz skupin gosto povezanih vozlišč, ki so med seboj povezane redko. Takšnim skupinam vozlišč pravimo skupnosti; na primer v družbenem omrežju skupnosti opisujejo osebe s podobnimi interesi [25], v omrežju citiranj pa znanstvene discipline [26]. Za odkrivanje skupin je bilo predlaganih veliko algoritmov [29, 30], pri čemer se večina osredotoča na skupine gosto povezanih vozlišč. Vozlišča pa se med seboj povezujejo tudi drugače; na primer moduli so skupine strukturno ekvivalentnih vozlišč [34]. Skupine vozlišč so redko raziskane v povezavi z zmanjševanjem omrežij. Salehi je s sodelavci [71] predlagal način zmanjševanja, primeren za omrežja z zelo izrazitimi skupnostmi. Maiya in Berger-Wolf [72] pa sta predlagala pristop, ki v zmanjšanem omrežju ohrani vozlišča iz vseh skupnosti v omrežju. Malo pa je znanega o tem, kako se skupine različnih vrst spreminjajo med zmanjševanjem.

V članku, vloženem v poglavju 5, analiziramo spreminjanje skupin vozlišč pri zmanjševanjem družbenih in informacijskih omrežjih. S pomočjo algoritma za odkrivanje skupin vozlišč [73, 74] v osnovnih in zmanjšanih omrežjih poiščemo skupine vozlišč, opisane s parametri kot so število vseh skupin v omrežju, število vozlišč v skupinah ter karakteristiko skupin, ki ponazarja, kako podobne so skupine skupnostim ali modulom. Izkaže se, da osnovna družbena omrežja vsebujejo gosto povezane skupine, podobne skupnostim, ki so med seboj povezane z manj povezavami. Po drugi strani so skupine v informacijskih omrežjih redke, bolj podobne modulom, med seboj povezane z več povezavami. Pri zmanjševanju z različnimi pristopi se skupine spremenijo. Tako v družbenih kot v informacijskih omrežjih postanejo gostejše, z manj povezavami med skupinami. Nekatere rezultate v članku smo potrdili vizualno in statistično. Statistična analiza ni bila narejena na naključnem vzorcu populacije, zato je uporabljene statistične teste in dobljene p-vrednosti potrebno primerno upoštevati z določeno toleranco ali dodatno preveriti.

3

Ocenjevanje učinkovitosti pristopov za zmanjševanje realnih omrežij

Many real-world networks are large, complex and thus hard to understand, analyze or visualize. Data about networks are not always complete, their structure may be hidden, or they may change quickly over time. Therefore, understanding how an incomplete system differs from a complete one is crucial. In this paper, we study the changes in networks submitted to simplification processes (i.e., reduction in size). We simplify 30 real-world networks using six simplification methods and analyze the similarity between the original and simplified networks based on the preservation of several properties, for example, degree distribution, clustering coefficient, betweenness centrality, density and degree mixing. We propose an approach for assessing the effectiveness of the simplification process to define the most appropriate size of simplified networks and to determine the method that preserves the most properties of original networks. The results reveal that the type and size of original networks do not affect the changes in the networks when submitted to simplification, whereas the size of simplified networks does. Moreover, we investigate the performance of simplification methods when the size of simplified networks is 10% that of the original networks. The findings show that sampling methods outperform merging ones, particularly random node selection based on degree and breadth-first sampling.

3.1 Introduction

Over the past decade, network analysis [1, 75] has proved to be a suitable tool for describing diverse systems, understanding their structure and analyzing their properties. However, the evolution of the Web and the capability of storing large amounts of data have caused the size of networked systems and thus their complexity to increase. The algorithms for analyzing and visualizing networks appear impractical for addressing very large systems. Therefore, different methods have been proposed for the simplification of complex networks.

Simplification is a process that reduces the size of a network by decreasing the number of nodes and links. The procedure is derived from graph theory (e.g., partitioning [76] and blockmodeling [77]) and was initially developed for compression and efficient graph storage [46, 47]. With the increasing complexity of networks, simplification methods also support clearer visualization [50, 51] and efficient analysis [40, 57]. In addition to these benefits, analyzing the changes undergone by networks under the effects of the sim-

plification process enables us to explore and explain the differences between complete (i.e., original) and incomplete (i.e., simplified) systems (e.g., when only partial insight into the structure of network is available).

Recently, network simplification has been extensively investigated from different perspectives. Some studies have concentrated on the simplification of specific networks, such as simplifying social networks based on stability and retention [78], sampling scale-free [52] or directed networks [53], estimating different properties under social network crawling [63], sampling large dynamic peer-to-peer networks with random walks [79] or simplifying flow networks by removing useless links [80]. Other studies have attempted to provide a sufficient fit to original networks and thus observe the changes in network properties under the effects of simplification, such as preserving the clustering coefficient [58], degree distribution [54], community structure [56], spectral properties [81] or network connectivity [55].

However, only a few studies have focused on comparing simplification methods and measuring their success. Leskovec et al. [40] observed properties of original and simplified networks submitted to several simplification methods and measured their success based on random walk similarity. Lee et al. [57] analyzed basic network properties under the effects of three simplification methods and revealed characteristic patterns of changes in properties. Hübler et al. [62] compared their simplification algorithm to existing ones by measuring the average distance of properties between original and simplified networks. Toivonen et al. [64] studied the compression of weighted networks and measured the method's efficiency according to the running time and cost of the compressed network representation. Doer and Blenn [63] tested the convergence of different properties under three traversal algorithms applied to a single large social network. The findings of the aforementioned analyses indicate that the performances of simplification methods vary; however, the common weakness of these studies is the small set of networks considered.

Despite the above-described efforts, several open questions remain concerning the simplification of complex networks, such as those regarding (Q1) how to evaluate the similarity between original and simplified network, (Q2) how small simplified networks should be and ultimately (Q3) what simplification method should be used. In this paper, we address these questions and propose an approach for assessing the effectiveness of the simplification process. We analyze 30 real-world networks of different size and origin under the effects of six different simplification methods. We compare the original and

simplified networks based on several network properties (e.g., degree distribution, clustering coefficient [22], betweenness centrality [67], degree mixing [65] and transitivity [66]) (Q1). The selection of these properties is supported by their common use in similar studies [40, 57]. Moreover, we propose a measure for determining the most appropriate size of simplified networks for preserving the observed properties (Q2) and for determining under which method the simplified networks fit the original ones most closely (Q3). We also study the impact of the original network size and type on the effectiveness of the simplification process.

The rest of the paper is structured as follows. Section 3.2 focuses on the simplification methods and real-world networks used in the study and describes the proposed measure. In section 3.3, we report and formally discuss the results of the analysis. Finally, section 3.4 concludes the paper and suggests directions for future research.

3.2 Methods and data

3.2.1 Simplification methods

Several authors have proposed a broad collection of simplification methods, which can be divided into two general classes. Those in the first class are sampling methods in which a simplified network is represented by a random sample of the original network (e.g., random node selection [48], random link selection [59], snowball sampling [82], random walk sampling [40] and forest fire [40]). Methods in the second class obtain simplified networks by merging nodes and links into supernodes and superlinks based on different characteristics, such as the distance between nodes (e.g., cluster-growing and box-tiling renormalization [37]), node and link attributes (e.g., link weights [83] and node attributes [84]) or community structure (e.g., balanced propagation and modularity optimization [38]).

In this study, we adopt four basic sampling methods (Fig. 3.1). Random node [48] (RN) and random link selection [59] (RL) create sampled networks with nodes or links selected uniformly at random. Simplified networks under random node selection based on degree [40] (RD) consist of randomly selected nodes, where the probability of selecting a node is proportional to the node's degree. In breadth-first sampling (BF), a random node with its broad neighborhood is selected into the sample using the breadth-first search strat-

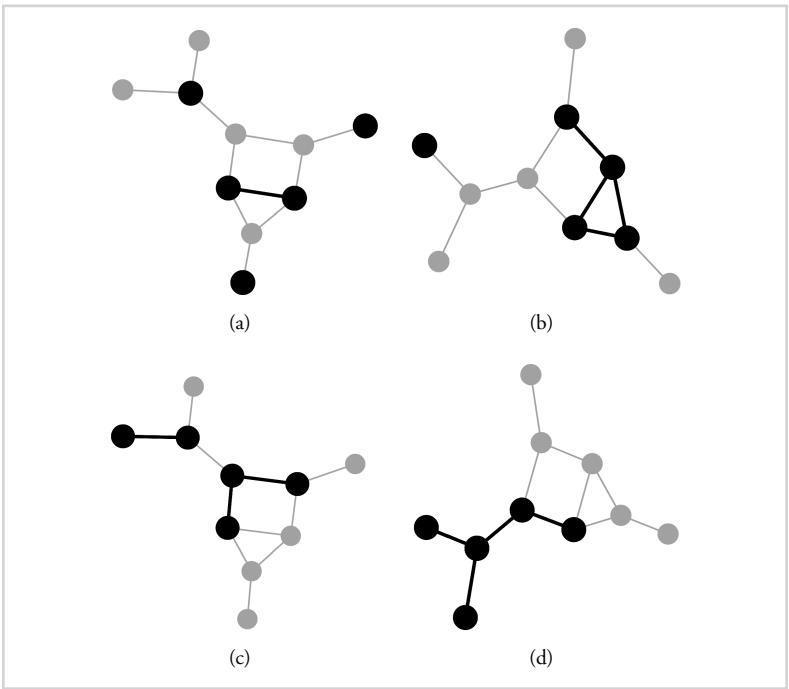


Figure 3.1

Sampling methods applied to a small sample network for $s = 0.5$. Black nodes represent simplified networks obtained with (a) selecting nodes uniformly at random (RN), (b) selecting nodes with probability proportional to degree (RD), (c) selecting links uniformly at random (RL) and (d) performing breadth-first search starting at a randomly selected node (BF). In the last method, BF ensures a connected network, whereas in other methods, this is not always the case.

egy. The main advantages of these methods are simplicity, and thus efficient implementation with low time complexity, and adjustability, which enables setting the size of the simplified network in advance.

Sampling methods outperform merging ones in terms of the advantages listed above. Still, we consider two methods from the merging class (Fig. 3.2). We use merging nodes based on community detection, where supernodes are identified by communities revealed by balanced propagation [85] (BP). We also employ cluster-growing renormalization [37] (CG), which incrementally grows supernodes from randomly selected seed nodes within a distance not larger than c (the nodes within one supernode are at most $2 \cdot c + 1$ steps apart). Both methods proved well in analyzing the invariance of network density under different renormalizations [38].

We define s as the number of nodes in the simplified network, measured

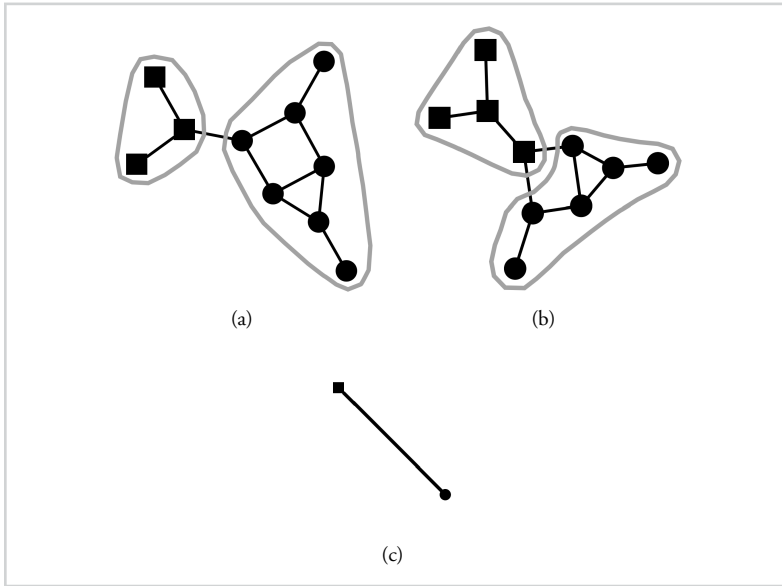


Figure 3.2

Merging methods applied to a small sample network. The shape of the nodes indicates (a) the nodes' community membership (BP) and (b) whether the nodes are at a distance less than 5 ($c = 2$) within one box (CG). Communities and boxes are marked by a gray contour. The simplified network (shown for both cases in (c)) is obtained by merging nodes inside one community or box into supernodes.

as the fraction of nodes in the original network. For sampling, we set the sizes of the simplified networks as varying from 1% to 50% of the original networks ($s = 0.01$ and $s = 0.05$-0.50 with a step size of 0.05). For BP, we set the parameters of the algorithm as suggested in [85]. With CG, we cannot control the size of the simplified network; still, we can change the distance between the nodes within one supernode. Therefore, the parameter c ranges from two to six, where smaller values indicate a smaller number of nodes within one supernode and thus a larger simplified network.

3.2.2 *Network data*

A diverse set of real-world systems is analyzed. We consider 30 networks of different origin (e.g., information, technological and social) and size (varying from a few thousand to a few hundred thousand nodes), listed in Table 3.1. Due to the large number of networks considered, a detailed description is omitted here.

For BP, CG and BF, all networks are considered to be undirected, al-

though some of them are directed. To avoid comparing networks of different complexity, we remove self-loops and multiple links from all networks for simplification via merging methods.

3.2.3 Assessment approach

To perform a fair and sound assessment, we first address the aforementioned questions concerning the comparison approach (Q1) and the size to which a certain network should be simplified (Q2). To address Q1, we select a set of local and global network properties to be observed. To address Q2, we introduce a simple measure that takes into account all of the selected properties and for each network calculates the simplified size that would best preserve the observed properties. The specific size of the simplified networks is then used in a further analysis to compare the selected simplification methods (Q3).

Comparing original and simplified networks

We compare networks based on eight fundamental global and local properties. The global properties are expressed by a single value for each network and include density (the ratio of existing links to all possible links), degree mixing (the tendency of nodes connecting to similar ones [65]) and transitivity (the number of closed triplets over the total number of triplets [66]). The local properties are described by a distribution for all nodes in the network and comprise degree, in-degree and out-degree (the number of neighbors of each node), local clustering coefficient (the proportion of connected neighbors of each node [22]) and betweenness centrality (the number of shortest paths between all nodes going through each node [67]).

For comparison, we define two similarity measures, one based on the selected global properties and one on the selected local properties. The global similarity measure is used to determine how correlated the global properties in the observed original networks and their simplified version are. The correlation is measured with Spearman's correlation coefficient ρ. ρ indicates the extent to which one variable decreases as another increases. In our analysis, we calculate ρ for each selected simplification method and each size of the simplified networks for all networks together.

The comparison based on the selected local properties is expressed using the Kolmogorov-Smirnov D-statistic (Kolmogorov-Smirnov test checks the null-hypothesis, i.e., that the distributions of two properties are the same; the

Table 3.1

Real-world networks (n and m correspond to the number of nodes and links, respectively).

Network	Type	n	m
High E. Particle Phys. [86]	Citation	27240	342437
High E. Phys. [87]		34546	421578
NBER US patents [88]		240548	561060
Citeseer publications [89]		384413	1764929
PGP web-of-trust [90]	Collaboration	10680	24340
High E. Phys. archive [91]		12008	237010
Astro Phys. archive [91]		18772	396160
Cond. Matters archive [91]		23133	186936
Computer science [92]		317080	1049866
Digg user reply [93]	Communication	30398	87627
Emails at *Enron* [94]		36692	367662
Facebook wall post [95]		46952	876993
Emails at EU res. inst. [91]		265214	420045
Amazon products 1 [94]	Co-purchase	334863	925872
Amazon products 2 [96]		403394	3387388
Flickr images metadata [97]	Co-occurence	105938	2316948
Oregon aut. systems [98]	Internet	22963	48436
Gnutella file sharing 1 [91]		36682	88328
Gnutella file sharing 2 [91]		62586	147829
Foldoc dictionary [99]	Information	13356	120238
Wikipedia votes [100]	On-line social	7115	103689
Brightkite friendship [101]		58228	214078
Epinions trust [102]		75879	508837
Slashdot friendship [103]		82168	948464
Wikipedia interactions [104]		186485	740397
Gowalla friendship [101]		196591	1900654
Broad-topic queries [105]	Web graph	6175	16150
google.com internal [106]		15763	171206
nd.edu domain [107]		325729	1497134
Baidu articles [108]		415641	3284387

D-statistic measures the distance between the observed distributions). The D-statistic for each network and its simplified version is calculated for each simplification method separately. The values for comparison based on ρ and the D-statistics are averaged over ten simplifications of each network, each simplification method and each size of the simplified networks.

The selection of properties and their relevance in assessing the effectiveness of network simplification greatly depends on the purpose of the simplification being performed. The selection of particular properties in this analysis is only supported by their common use in similar studies (e.g., [40, 57]) and serves to demonstrate the effectiveness of the proposed approach. Note that comparing networks based on other sets of properties may lead to different results.

In the literature, we can find studies that have performed similar comparisons to a limited extent. In [52] the authors proved that RN does not preserve the degree distribution of scale-free networks. Moreover, RN and RL sampling are biased toward nodes with high degrees, which affects the degree distribution [40]. However, Lee et al. [57] proved that RN and RL overestimate the degree and betweenness centrality exponent, whereas both methods retain the assortativity of original networks. Merging methods decreases the density [38], but the relationship between density and network size remains invariant after simplification.

Determining simplified network sizes

To determine the size to which a specific network can be decreased while preserving most of the observed properties, the following approach is used. For each simplification method and each global and local property, we rank sizes with respect to ρ and the D-statistic, respectively. The network size that best fits a specific property receives rank 0, the next best one receives rank 1 and so on. Next, we sum the ranks for each size and divide the sum by the greatest possible sum of ranks to normalize the result to the interval $[0, 1]$. Thus, the measure A is defined as

$$A = \frac{1}{(n_s - 1) \cdot n_p} \sum_{i=1}^{n_p} r_i, \tag{3.1}$$

where n_s denotes the number of different sizes, n_p denotes the number of properties, i indexes the properties (the order is not important) and r_i is the

Table 3.2

An illustrative example of the assessment approach. (left) The results of a comparison between simplified and original networks based on global properties. (right) The results after ranking sizes for each property. P*i* denotes properties and S*i* sizes of simplified networks.

	P1	P2	P3
S1	0.84	0.69	0.75
S2	0.88	0.89	0.87
S3	0.90	0.92	0.89
S4	0.96	0.95	0.88
S5	0.91	0.94	0.92
S6	0.93	0.96	0.90

	P1	P2	P3	Sum	A
S1	5	5	5	15	1.000
S2	4	4	4	12	0.800
S3	3	3	2	8	0.533
S4	0	1	3	4	0.267
S5	2	2	0	4	0.267
S6	1	0	1	2	0.133

rank of the i−th property. A is thus the normalized total rank assigned to a specific size by the observed properties.

Table 3.2 shows an example of the measure A calculated by comparing six different sizes for a simplified network, taking into account the measure ρ a specific size receives for each of the three observed global properties. In this example, the most appropriate size for preserving global properties is S6.

Comparing simplification methods

Finally, we compare the different methods for a given size of a simplified network. We rank the methods and measure their effectiveness using a modified version of the measure A described in the previous subsection:

$$A = \frac{1}{(n_m - 1) \cdot n_p} \sum_{i=1}^{n_p} r_i, \tag{3.2}$$

where n_m is the number of different methods.

With the described measure, we regard all properties as equally important. Still, depending on the purpose of the simplified networks considered and the method by which those networks are analyzed, one property can be more essential than another. With respect to importance, we can assign weights w to the properties; thus, the measure A becomes

$$A_w = \frac{1}{n_m \cdot \sum_{i=1}^{n_p} w_i} \sum_{i=1}^{n_p} r_i \cdot w_i, \tag{3.3}$$

where w is the vector of weights and thus w_i denotes the weight of property i.

For simplicity, we omit the analysis performed based on the measure A_w and thus assume all properties are equally important.

3.3 Analysis and discussion

The analysis consists of two stages. First, we determine the size of the simplified networks that ensure adequate preservation of the observed properties. Second, we compare the effectiveness of different methods for a specific size of the simplified networks.

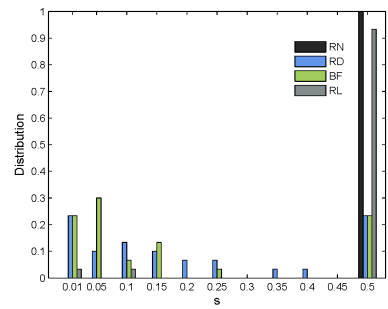
3.3.1 Effectiveness of the simplification process with respect to the size of the simplified networks

First, we analyze the effect of simplified network size on the effectiveness of the simplification process. As expected, the results reveal that in the majority of cases, the largest simplified networks ($c = 2$ for CG and $s = 0.5$ for sampling methods) are more similar to the original networks and thus better fit the original networks' properties. However, the main goal of the simplification is to sufficiently reduce large networks to allow for easier analysis and understanding, which is achieved when the simplified networks are smaller. Therefore, we define the best size as the local minimum of A achieved at the smallest simplified network size (we assume that $A = 1$ for $s = 0$ and take the global minimum if it is also local).

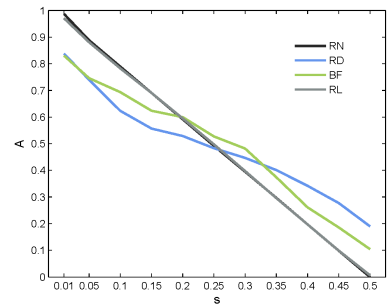
Analysis of the sampling methods

The analysis of the sampling methods reveals a high level of diversity in their effectiveness (Fig. 3.3 and Table 3.3). Fig. 3.3(a) shows that under simplification methods RN and RL, local properties are best preserved for the largest size of the simplified networks ($s = 0.5$). In contrast, RD and BF perform best for smaller sizes, between $s = 0.01$ and $s = 0.15$, for the majority of the networks (i.e., the local minimum of A is around these values for most of the networks).

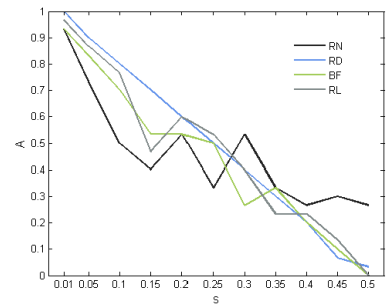
Fig. 3.3(b) and 3.3(c) shows the average A over all networks for the local and global properties, respectively. For the former, all methods behave in a similar manner. In particular, the best fit of local properties is reached for larger simplified networks; still, RD and BF show some deviation, indicating that for several networks smaller sizes also provide good fits. For the global



(a)



(b)



(c)

Figure 3.3

The results for the sampling methods. (a) Portion of networks with the best size equal to s. (b) Distance between the original and simplified networks (average A over all networks) based on the local properties. (c) Distance between original and simplified networks based on the global properties.

Table 3.3

The best sizes c or s for the preservation of the local properties with corresponding A.

Network	CG	RD	BF
High E. Particle Phys.	2 (0.00)	0.20 (0.18)	0.05 (0.38)
High E. Phys.	2 (0.08)	0.25 (0.14)	0.10 (0.44)
NBER US patents	2 (0.25)	0.35 (0.20)	0.01 (0.44)
Citeseer publications	2 (0.00)	0.20 (0.08)	0.01 (0.58)
PGP web-of-trust	4 (0.33)	0.25 (0.40)	0.05 (0.77)
High E. Phys. archive	2 (0.00)	0.05 (0.73)	0.05 (0.83)
Astro Phys. archive	2 (0.00)	0.50 (0.17)	0.05 (0.50)
Cond. Matters archive	2 (0.00)	0.50 (0.13)	0.05 (0.70)
Computer science	2 (0.17)	0.50 (0.00)	0.50 (0.00)
Digg user reply	2 (0.25)	0.10 (0.20)	0.05 (0.28)
Emails at *Enron*	2 (0.00)	0.01 (0.57)	0.50 (0.00)
Facebook wall post	2 (0.00)	0.10 (0.18)	0.01 (0.40)
Emails at EU res. inst.	2 (0.00)	0.50 (0.00)	0.15 (0.60)
Amazon products 1	4 (0.33)	0.50 (0.00)	0.50 (0.00)
Amazon products 2	2 (0.00)	0.50 (0.00)	0.50 (0.02)
Flickr images metadata	3 (0.33)	0.01 (0.80)	0.25 (0.37)
Oregon aut. systems	2 (0.17)	0.40 (0.23)	0.01 (0.93)
Gnutella file sharing 1	5 (0.58)	0.15 (0.34)	0.15 (0.34)
Gnutella file sharing 2	5 (0.42)	0.15 (0.34)	0.10 (0.36)
Foldoc dictionary	2 (0.17)	0.50 (0.00)	0.50 (0.02)
Wikipedia votes	2 (0.00)	0.01 (0.26)	0.01 (0.76)
Brightkite friendship	2 (0.00)	0.05 (0.37)	0.05 (0.83)
Epinions trust	2 (0.00)	0.01 (0.94)	0.01 (0.70)
Slashdot friendship	2 (0.17)	0.01 (0.30)	0.01 (0.58)
Wikipedia interactions	4 (0.47)	0.01 (0.42)	0.05 (0.44)
Gowalla friendship	2 (0.00)	0.05 (0.33)	0.05 (0.03)
Broad-topic queries	4 (0.33)	0.10 (0.26)	0.15 (0.34)
google.com internal	2 (0.00)	0.15 (0.34)	0.15 (0.36)
nd.edu domain	4 (0.08)	0.01 (0.48)	0.50 (0.32)
Baidu articles	2 (0.00)	0.10 (0.22)	0.05 (0.50)

properties, RN and RL show similar behavior again because the best preservation is achieved on smaller simplified networks ($s = 0.15$). For BF and RD, the local and global minima are reached for larger simplified networks.

Table 3.4 shows the best sizes of simplified networks for the preservation of each network property. RN and RL perform similarly because both provide better preservation of local properties for the largest simplified networks. On the other hand, for RD, degree is best preserved for smaller networks, whereas for medium-sized networks, out-degree and clustering are best preserved. For BF, distributions of degree, out-degree and in-degree change the least for $s = 0.01, 0.15$. However, the methods behave in a different manner when preserving global properties. Only RN preserves density and degree mixing well on smaller simplified networks, whereas RD, BF and RL work best for $s = 0.5$.

Finally, we analyze how the preservation of local properties depends on the size and type of the original networks (Table 3.3). We omit the results for RN and RL because in all cases except two, the best size is $s = 0.5$. In contrast, the effectiveness of RD is partially correlated to the original network size because medium-sized networks ($n = 50000 - 200000$) are best preserved for smaller simplified network sizes ($s = 0.01 - 0.1$), whereas large networks ($n = 200000 - 500000$) are best preserved for larger values of s. However, as indicated by the dependence on network type, the local properties of online social networks and Web graphs are best preserved for smaller sizes $s = 0.01 - 0.15$, whereas the local properties of citation and co-purchase networks are best preserved for $s = 0.25 - 0.35$. All differences are statistically significant ($p < 0.05$, one-way ANOVA), which rejects the null hypothesis that there is no dependence between the effectiveness of property preservation and network type. For both RD and BF, only the properties of co-purchase and information networks are best preserved for $s = 0.5$. The results reveal no statistically significant influence of network size or type on the performance of BF.

Analysis of the merging methods

The analysis of CG proves that the local network properties are best preserved when $c = 2$ for 22 out of 30 networks (Fig. 3.4(a) and Table 3.3). Fig. 3.4(b) shows the average A over all networks based on the local and global properties. The local properties are best fitted for larger simplified networks ($c = 2$), whereas for $c = 3, 4$ the simplified networks best fit the global properties of

Table 3.4

The best sizes c or s for the preservation of local network properties with corresponding A, and ρ for the global properties.

Property	CG	RN	RD	BF	RL
Degree	2 (0.04)	0.50 (0.00)	0.15 (0.52)	0.15 (0.61)	0.50 (0.00)
In degree	-	0.50 (0.00)	0.50 (0.26)	0.01 (0.75)	0.50 (0.02)
Out degree	-	0.50 (0.00)	0.30 (0.43)	0.01 (0.61)	0.50 (0.01)
Clustering	2 (0.25)	0.50 (0.00)	0.25 (0.44)	0.50 (0.10)	0.50 (0.00)
Betweenness	2 (0.00)	0.50 (0.00)	0.50 (0.08)	0.50 (0.05)	0.50 (0.01)
Density	2 (0.89)	0.10 (0.97)	0.45 (0.95)	0.50 (0.95)	0.50 (0.91)
Degree mixing	3 (0.34)	0.35 (0.66)	0.50 (0.77)	0.50 (0.97)	0.50 (0.63)
Transitivity	4 (0.36)	0.50 (0.99)	0.50 (0.99)	0.50 (0.99)	0.50 (0.83)

the original networks.

Table 3.4 shows the results obtained for the preservation of each property. Most of the properties are best preserved for larger simplified networks ($c = 2$), with the exception of degree mixing and transitivity, where $c = 5$ and $c = 6$, respectively.

The best size for preserving local network properties (Table 3.3) does not depend on the original network size or type (i.e., the differences in property preservation, which would depend on the size and type of the original networks, are not statistically significant). Still, if we divide the networks roughly by type, i.e., information, social and technological, the correlation between the type and the effectiveness becomes statistically significant (i.e., the null hypothesis that there are no differences in property preservation, which would depend on network type, is rejected, with $p < 0.05$, one-way ANOVA). Thus, the local properties of social networks are best preserved for $c = 2$, in contrast to the case of technological networks, for which $c > 2$.

Discussion

The findings of the first part of the study confirm the negative correlation between the size of the simplified networks and their similarity to the original networks because larger simplified networks are more similar to the original ones in most cases. The latter has also been proved by other studies, for example, [57]. RD and BF are more effective for smaller simplified networks, which is consistent with the findings of other authors. Particularly, Doerr and Blenn [63] revealed a solid estimate of an original network for $s = 0.2 - 0.3$

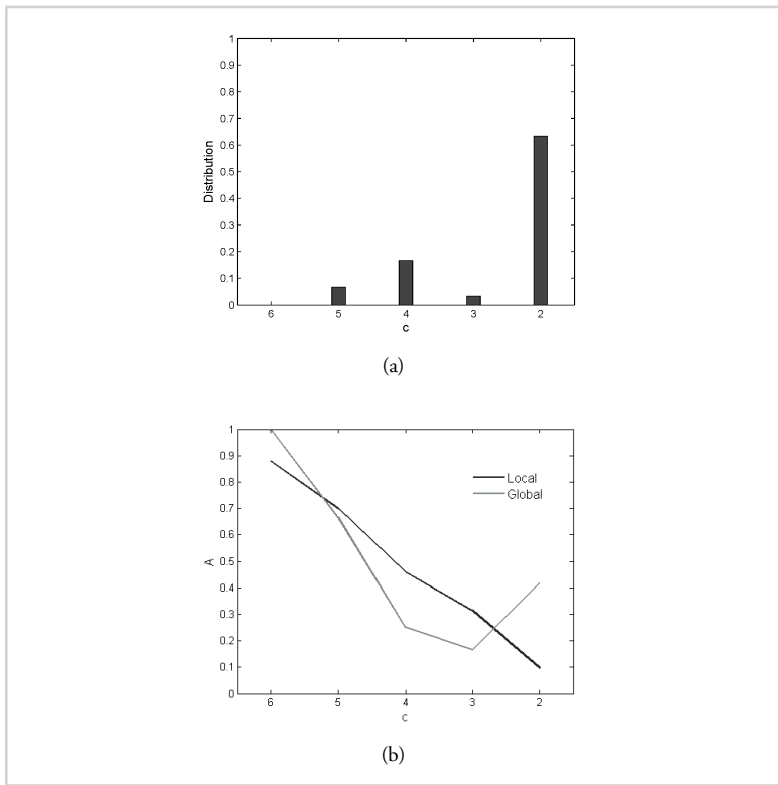


Figure 3.4

The results for cluster-growing simplification. (a) Portion of networks with the best size equal to c. (b) Distance between the original and simplified networks (A for the global and average A over all networks for the local properties) as a function of c.

Table 3.5

The best, second-best and worst methods for the preservation of local network properties with corresponding A, and ρ for the global properties.

Property	Best	Second-best	Worst
Degree	BF (0.25)	RD (0.26)	RL (0.84)
In degree	RD/BF (0.26)	RL (0.70)	RN (0.77)
Out degree	RD (0.32)	BF (0.33)	RL (0.70)
Clustering	RD (0.30)	BF (0.35)	RL (0.81)
Betweenness	BF (0.21)	RD (0.27)	BP (0.75)
Density	RN (0.96)	BF (0.91)	BP (0.76)
Degree mixing	BF (0.92)	RN (0.62)	BP (0.21)
Transitivity	RN (0.94)	RD (0.92)	CG (0.22)

and $s = 0.1$ in the case of preserving average node degree and the power-law degree exponent, respectively. In addition, Leskovec and Faloutsos [40] obtained a good fit for original networks under several sampling methods for $s = 0.15$. Thus, our results advance those reported in these studies and reveal distinctions in the extent of property preservation among different types and sizes of networks, which are the most obvious for RD.

3.3.2 Comparison of the effectiveness of the simplification methods

In the second part of our study, we compare the performance of different simplification methods. We focus on size $c = 2$ for CG and $s = 0.1$ for sampling methods for two reasons. First, we select $s = 0.1$ as the middle size among the best sizes determined in the first part of the study. Second, $s = 0.1$ is suitable for the comparison of BP and CG, for which the mean sizes of simplified networks are $s = 0.03$ and $s = 0.12$, respectively.

Analysis

First, we determine the best method for preserving a specific property (Table 3.5). Global properties are best preserved under RN and BF, whereas merging methods provide the worst preservation. Fig. 3.5 compares the best, second-best and worst methods with respect to all global properties. For local properties, BF and RD perform the best, particularly BF for the degree and betweenness centrality, whereas RD performs best for the out-degree and clustering. However, RL proves to be the worst method because it preserves the degree, out-degree and clustering to the lowest extent. Examples of local

Table 3.6

The best, second-best and worst methods for preserving local properties of networks with corresponding A.

Network	Best	Second-best	Worst
High E. Particle Phys.	RD (0.10)	BF (0.17)	RL (0.97)
High E. Phys.	BF (0.07)	RD (0.20)	RL (0.96)
NBER US patents	BF (0.07)	BP (0.13)	RN/RL (0.80)
Citeseer publications	RD (0.07)	BF (0.20)	RL (0.93)
PGP web-of-trust	CG (0.13)	BP (0.20)	RN (0.93)
High E. Phys. archive	RD (0.07)	BF (0.20)	RL (0.93)
Astro Phys. archive	RD (0.07)	BF (0.13)	RL (1.00)
Cond. Matters archive	BF (0.00)	RD (0.20)	RL (1.00)
Computer science	BF (0.07)	RD (0.27)	RL (1.00)
Digg user reply	RD (0.17)	CG/BP (0.33)	RL/RN (0.60)
Emails at *Enron*	BP (0.27)	RD (0.33)	RN (0.73)
Facebook wall post	RD (0.07)	BP (0.17)	RL (1.00)
Emails at EU res. inst.	RL (0.13)	BP (0.20)	RD (0.73)
Amazon products 1	BF (0.00)	BP (0.27)	RL (1.00)
Amazon products 2	BF (0.03)	CG (0.10)	RL (1.00)
Flickr images metadata	RD/BF (0.33)	RN (0.47)	RL (0.73)
Oregon aut. systems	RD (0.07)	BP (0.20)	RN (0.80)
Gnutella file sharing 1	BF (0.13)	BP (0.30)	RL (0.70)
Gnutella file sharing 2	BF (0.13)	BP (0.30)	RL (0.70)
Foldoc dictionary	BF (0.03)	CG/BP (0.13)	RL (1.00)
Wikipedia votes	BP (0.13)	RN (0.27)	BF (0.60)
Brightkite friendship	RD (0.13)	BP (0.20)	RL (0.93)
Epinions trust	BP (0.03)	RL (0.17)	BF (0.87)
Slashdot friendship	BP (0.07)	RD (0.23)	RL (0.83)
Wikipedia interactions	BP (0.07)	BF (0.33)	RN (0.40)
Gowalla friendship	RD (0.07)	BP (0.27)	RL (1.00)
Broad-topic queries	RD (0.07)	BP (0.27)	RN (0.73)
google.com internal	RD (0.10)	BF (0.17)	RL (0.93)
nd.edu domain	CG (0.07)	BP/BF (0.13)	RL/RN (0.80)
Baidu articles	RD (0.00)	BP (0.27)	RL (0.90)

property preservation for the analyzed networks are presented in Fig. 3.6 – 3.7.

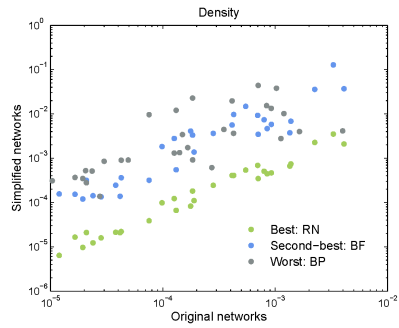
For a complete assessment of the effectiveness of the simplification methods, we compare the performance of the methods for each network based on the preservation of local properties. Results are represented in Table 3.6. For 23 networks, the best methods are RD and BF. The analysis reveals a dependence between network type and method effectiveness because BP performs the best for on-line social networks and BF performs the best for Internet and co-purchase networks. The differences among the network types are statistically significant ($p < 0.05$, one-way ANOVA). For the second-best methods, the distinctions are less evident. Still, BP proves to be effective for other types of networks (Internet, communication networks, Web graphs). The worst method for preserving local properties is RL (for 22 networks), followed by RN (for 8 networks). On the other hand, BF is the worst with respect to only two on-line social networks. The results also prove the statistically significant dependence (i.e., reject the null hypothesis that there are no dependencies between the network size and the effectiveness of the simplification methods, with $p < 0.05$, one-way ANOVA) between the worst method and network size. For smaller networks ($n < 50000$), the worst method for preserving local properties is RL, whereas for larger ones, the worst method is RN.

Discussion

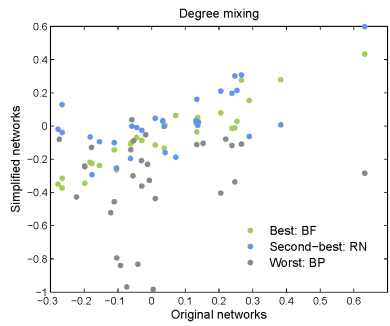
The results of the second part of the study reveal several distinctions in the behavior of the simplification methods. RD and BF proved the best for preserving the local properties of networks, whereas for global properties, RN outperforms the other methods. However, RL and merging methods show the worst performance. These findings are consistent with the results of the study reported in [40], where RD had a better performance than RN and RL (other methods are not considered in the aforementioned study).

In addition to comparing the methods for $s = 0.1$, we also compare them for larger simplified networks ($s = 0.5$). The results are not presented because there are no significant changes in the results (i.e., the same methods are the best and the worst for $s = 0.1$).

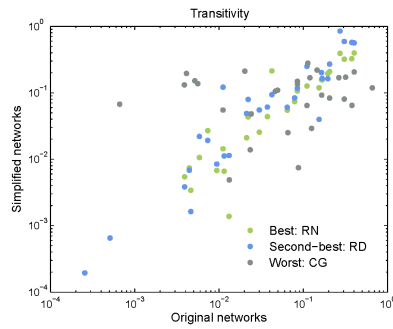
In addition, we observe how the size of the largest weakly connected component (LWCC) changes under simplification to explain the differences in the methods' performance. The LWCC of the original networks, on average, consists of 59% of all nodes. The size of the LWCC of the simplified networks



(a)



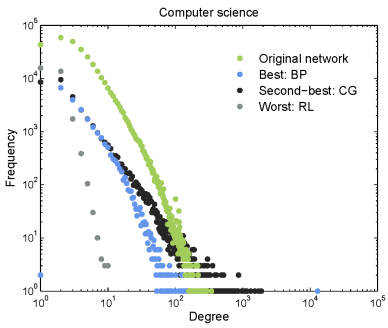
(b)



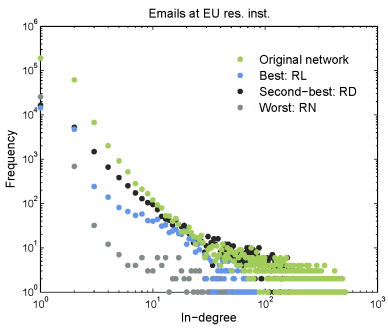
(c)

Figure 3.5

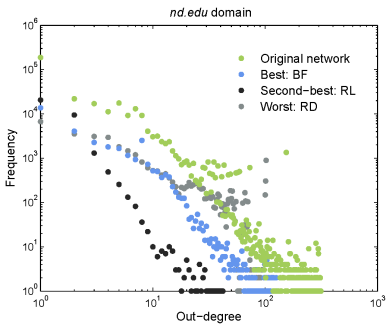
Relationship between the global properties of the original and the simplified networks for the best, second-best and worst method. (a) Density. (b) Degree mixing. (c) Transitivity.



(a)



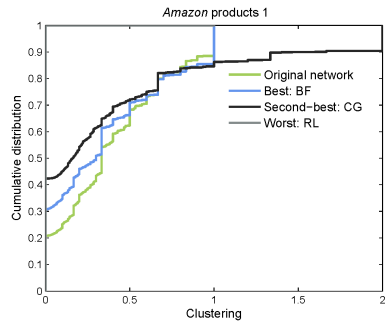
(b)



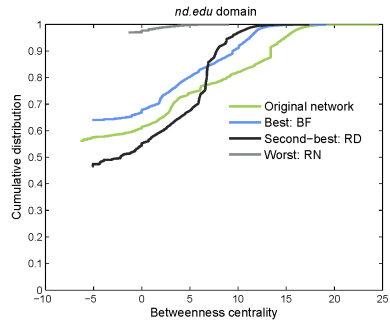
(c)

Figure 3.6

Examples of comparison of the local properties for the original and simplified networks with the best, second-best and worst methods. (a) Degree distribution. (b) In-degree distribution. (c) Out-degree distribution.



(a)



(b)

Figure 3.7

Examples of comparison of the local properties for the original and simplified networks with the best, second-best and worst methods. (a) Cumulative distribution of clustering. (b) Cumulative distribution of of betweenness centrality.

under all methods depends strongly on the simplified network size (i.e., the size of the LWCC of the smallest simplified networks is the smallest). However, RN and RL show similar performance because the sizes of the LWCC for both methods vary from 1% for $s = 0.1$ to 40% for $s = 0.5$. Still, RL produces the most disconnected components. In contrast, simplification via RD and BF produces networks with a clearly larger LWCC because the sizes vary from 25% for $s = 0.01$ to 60% for $s = 0.5$. Therefore, networks simplified by RD and BF feature a larger LWCC and smaller number of components, which is more similar to the characteristics of the original networks. Based on this finding, the predominance of RD and BF over RN and RL can be confirmed.

3.4 Conclusions

Network simplification is an adequate tool for studying large networks for several reasons. In addition to the obvious advantages, including faster analysis and more efficient visualization, the simplification can significantly improve the understanding of large networks. For example, data regarding the systems described by networks can often be missing or incomplete, and thus, networks can be considered a sampled variety of the original systems (e.g., identifying Internet map [109, 110]). For this reason, understanding how similar the original and sampled system are is essential.

This study addressed three aspects of real-world network simplification. First, we focused on a comparison of original and simplified networks. Second, we determined what size of simplified network most adequately fits the properties of the original networks. Finally, we compared the effectiveness of several simplification methods. We analyzed six simplification methods with respect to 30 real-world networks and compared the simplified and original networks based on several properties, including degree, in-degree, out-degree and betweenness centrality distribution, clustering coefficient, density, degree mixing and transitivity.

The results show that the goodness of property preservation depends on the size of the simplified networks. Larger simplified networks fit original networks better; nevertheless, properties are adequately preserved for smaller sizes close to 10% the size of the original networks, especially for random node selection based on degree and breadth-first sampling. Thus, the decision regarding how small a simplified network should be depends on the size of

the original network and the purpose of the simplified network. If we can simplify a network by 50%, we can provide for the best fit of the original network properties. However, if the network is large, 50% of the original size is not a sufficient reduction. In that case, 10% of the original network size allows for the adequate preservation of important properties. Furthermore, the findings of this study reveal that random node selection based on degree and breadth-first sampling are the best methods, whereas merging methods performed the worst.

Future work will mainly focus on other characteristics that affect the effectiveness of the simplification process. Moreover, instead of focusing solely on similarities, we will analyze typical distinctions between original and simplified networks. Furthermore, other ways for comparing simplified networks with original for their similarity could also be considered, for example comparing the backbones of networks [111], their community structure [112] or density of edges in subnetworks [113]. Based on this and future studies, a wide range of principles underlying the simplification of real-world networks could be extracted. The application of such principles should allow for the determination of the most suitable simplification method for specific networks, which would allow for more efficient simplification and a better understanding of large real-world networks.

Acknowledgment

The work has been supported by the Slovene Research Agency *ARRS* within the research program P2-0359.

4

Samopodobnost gostote realnih omrežij

Despite their diverse origin, networks of large real-world systems reveal a number of common properties including small-world phenomena, scale-free degree distributions and modularity. Recently, network self-similarity as a natural outcome of the evolution of real-world systems has also attracted much attention within the physics literature. Here we investigate the scaling of density in complex networks under two classical box-covering renormalizations—network coarse-graining—and also different community-based renormalizations. The analysis on over 50 real-world networks reveals a power-law scaling of network density and size under adequate renormalization technique, yet irrespective of network type and origin. The results thus advance a recent discovery of a universal scaling of density among different real-world networks [24] and imply an existence of a scale-free density also within—among different self-similar scales of—complex real-world networks. The latter further improves the comprehension of self-similar structure in large real-world networks with several possible applications.

4.1 Introduction

The study of complex real-world networks and underlying systems has erupted in recent years in various fields of science. Due to their simple and intelligible form, networks enable representation of diverse systems of complex interactions and provide for their common investigation. Thus, several fundamental properties of large real-world networks have been revealed in the past decade. These include small-world phenomena [22], scale-free degree distributions [114, 115], network clustering [22, 116] and robustness [117, 118], degree mixing [36, 65], community and hierarchical structure [27, 119], network motifs [120] and other [121] (for reviews see [17, 122]). More recently, network self-similarity as an inherent property behind the evolution of real-world systems has also attracted much of attention within the physics community [28, 37, 68–70, 123].

Network self-similarity is commonly considered alongside the concept of fractal networks [28, 124]. Fractality is a property of a geometric object that it is exactly or approximately similar to a part of itself [125]. Nevertheless, classical theory of self-similarity requires a power-law scaling between the system size and its parts under some renormalization [126, 127]. The latter is an iterative process where a system is coarse-grained into smaller replicas, thus its essential structural features are preserved [28, 128] (Fig. 4.1). Hence,

fractal or self-similar networks commonly refer only to a self-similar scaling exponent in the afore mentioned power-law relation [28, 69, 129, 130]. However, network self-similarity is also investigated in the context of other network properties [68–70, 123, 131] under various renormalization techniques [28, 128, 132, 133].[1] (Note that fractal scaling laws observed in real-world networks do not necessarily imply a self-similar network [134].)

Guimerà et al. [68] have first observed self-similar community size distributions in a network of human communications. Furthermore, Song et al. [28, 129] have proposed an adequate renormalization technique (Fig. 4.1) to expose the origin of self-similar fractal scaling in web, collaboration and different biological networks. The latter in fact gives rise to degree disassortativity [129] and resilience to diseases [135], commonly observed for these networks. Still, such scaling cannot coexist with a small-world network topology [136, 137]. Self-similarity has also been considered as a scale-invariance of degree distribution [69, 130] or maximum degree [70, 138] under network renormalization, while Itzkovitz et al. [123] have revealed self-dissimilarity in a motif structure for different biological and technological networks. Authors have also considered network self-similarity in the context of different dynamical processes including percolation [139] and synchronization [140].

Despite the above efforts, there is yet little evidence whether self-similarity exists only in certain networks and which properties are indeed invariant throughout different network scales. We thus here investigate the scaling of density—defined as the number of links to all possible links—with respect to network size under five renormalization techniques borrowed from the field of fractal networks [28, 129] and community detection literature [29, 30]. Analysis on over 50 real-world networks of diverse origin reveals a self-similar power-law scaling of network density and size (under suitable renormalization). The latter advances a recent work of Laurienti et al. [24] who have observed a universal scaling of density among different real-world networks, while Leskovec et al. [87, 91] have also found similar densification laws in evolving networks. The results thus imply an existence of a scale-free density not only among, but also within—among different self-similar scales of—complex networks irrespective of their type and the underlying domain. Hence, under adequate renormalization self-similar real-world networks nei-

[1] Throughout the paper we refer to network self-similarity in a general sense.

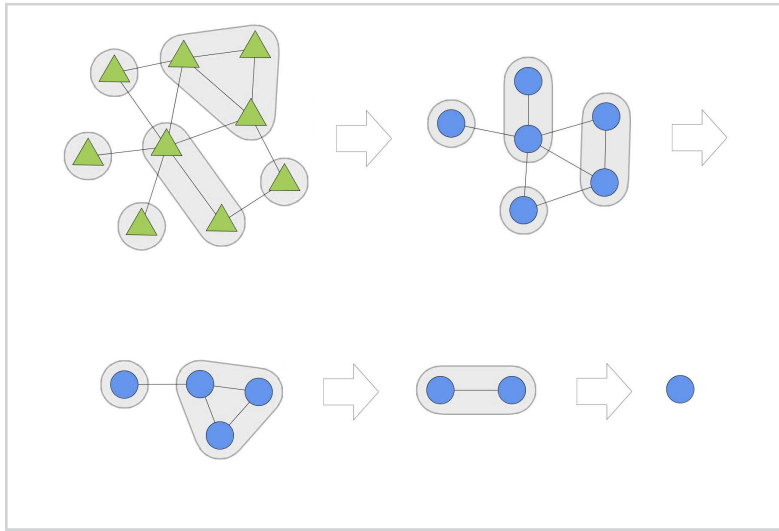


Figure 4.1

Network renormalization—system coarse-graining—technique [28, 128] applied to a small example network. At each step, the network is covered with boxes that are replaced by super-nodes. The latter are linked when a corresponding link also exists in the (original) network. The process then repeats until only a single node remains or multiple nodes in the case of a disconnected network. (Here the network is randomly tiled with boxes of nodes at a distance smaller than 2.)

ther get denser nor sparser with respect to their size, whereas characteristic network topology is also largely retained throughout the renormalization.

The rest of the paper is structured as follows. Section 4.2 introduces different renormalization techniques and real-world network data adopted in the research. Empirical analysis with formal discussion on real-world and random networks is presented in section 4.3, while section 4.4 gives final conclusions and discusses future work.

4.2 *Techniques and network data*

Self-similarity is primarily studied under the framework of network renormalization [28, 69]. As already discussed, renormalization is an iterative coarse-graining technique, where the original network is covered with boxes, thus each node belongs to exactly one box [28, 128] (Fig. 4.1). Boxes are then replaced by super-nodes that are linked when a corresponding link also exists in the (original) network. The entire process repeats until no links remain and the number of nodes equals to the number of connected components.

While there exists a number of different box-covering approaches, not all

of them are able to reveal self-similar scales in complex networks. Thus, we employ techniques that have already proven useful for exposing self-similarity in various real-world networks [28, 68–70]. In particular, we adopt methods commonly used in analysis of fractal networks and as well as different community detection algorithms.

Fractal network structure is mainly explored under two general classes of renormalization techniques, namely, node coloring and network burning approaches [28, 132] (for reviews see [37, 130]). In the former, box-covering is mapped to a node coloring problem [141, 142], whereas, in the latter, boxes are grown around a randomly selected seed node. Although there exist several efficient algorithms for node coloring [142, 143], network burning methods offer some distinct advantages [132]. Different authors have proposed a wide range of alternative network coarse-graining techniques including methods based on connectivity patterns [123], skeleton of the network [128], link-covering [39] and other [69, 133, 140, 144].

For the purpose of this research, we adopt two classical network burning approaches. First, box-tiling method, randomly tiles the network with boxes of nodes that are at a distance smaller than l_B [28, 129] (Fig. 4.1). Second, cluster-growing method, incrementally grows boxes from randomly selected seed nodes within a distance not larger than r_B [132, 144]. Hence, for random configurations, $l_B = 2 \cdot r_B + 1$ [132]. Box-tilling method allows for somewhat easier analytical consideration, whereas cluster-growing approach enables more efficient implementation. For the analysis in section 4.3, we set l_B to 3 and r_B to 2 with respect to network small-worlds [22]. Note that the latter extends the definition of an egonet [145, 146]—a subnetwork inferred by a central ego node and its neighbors—which can be seen as a local signature of the respective node.

We further adopt several algorithms drawn from community detection literature (for reviews see [29, 30]). Here boxes are identified by communities [27]—groups of nodes densely connected within and only loosely connected between—revealed with selected algorithm, whereas network coarse-graining procedure is else identical as above. Community detection has already been successfully employed to reveal self-similarity in real-world networks [68]. Recent work also implies an existence of community structures on various scales of complex real-world networks [147, 148]. Hence, community detection appears to be an adequate alternative to classical box-covering renormalization techniques.

Due to generality, we consider three diverse community detection algorithms. First, we adopt balanced propagation [85] as an example of a highly scalable state-of-the-art algorithm. The approach is based on the label propagation principle of Raghavan et al. [149], while node balancers are introduced to improve the stability of the algorithm (stability parameter is set to 1/4). Next, we employ a fast hierarchical optimization of modularity Q [150] proposed by Clauset et al. [151] as one of most widely used approaches in the past literature [30]. However, due to many limitations of the measure of modularity Q, high values of Q cannot be regarded as an indication of network community structure [152, 153]. Last, we also consider a spectral algorithm of Newman [154] as a representative of a partitioning approach with origins in classical graph theory [155]. The algorithm reveals communities by extracting the leading eigenvector of network modularity matrix using a power method.

Analysis in section 4.3 is conducted on 55 real-world networks that are often analyzed in complex network literature (Tab. 4.1 – 4.2) and also on random graphs á la Erdős-Rényi [156]. The real-world networks range between tens of nodes and tens of millions of links; and include different social—classical, on-line, collaboration etc.; information—web graphs, citation, communication etc.; technological—Internet, software, transportation etc.; biological—protein, genetic and neural; and other networks. Due to the large number of networks considered, detailed description is omitted. Still, networks were carefully chosen thus to represent a relatively diverse set of real-world systems including most types of networks commonly analyzed in the literature. For simplicity, all networks are considered as simple undirected graphs and reduced to largest connected components.

4.3 *Analysis and discussion*

In the following section we analyze self-similar scaling of density in real-world networks of moderate size (section 4.3.1) and different Erdős-Rényi random graphs (section 4.3.2); whereas in section 4.3.3 we further consider self-similarity of five larger real-world networks with at least a million links.

4.3.1 Real-world networks

The algorithms were first applied to 50 real-world networks (Tab. 4.1 – 4.2). According to the number of nodes n and density d from original and reduced networks we examine the density scaling with respect to network size. In

Table 4.1

Real-world networks. (n and m correspond to the number of nodes and links, respectively.)

Network	Type	n	m
Zachary's karate club [10]	Social	34	78
Lusseau's dolphins [157]		62	159
Comp. sci. PhD students [158]		1025	1043
Facebook friendships [159]	On-line social	324	2218
Wikipedia who-votes-who [100]		7066	100736
Slovenian comp. science [159]	Collaboration	239	568
Krebs's Internet industry [158]		219	630
Complex networks science [154]		379	914
Paul Erdős collaborations [158]		446	1413
Comput. Geometry archive [160]		3621	9461
General Relativity archive [87]		4158	13422
PGP web-of-trust [90]		10680	24316
Astro Physics archive [91]		17903	196972
US political books [161]	Co-purchase	105	441
amazon.com domain [34]	Web graph	2879	3886
epa.gov domain [158]		4253	8897
Broad-topic queries [105]		5925	15770
US political blogs [161]		1222	16714
Graph Drawing proceedings [158]	Citation	249	635
Stanley Milgram citations [158]		233	994
H. Small & B. Griffith citations [158]		1024	4916
Scientometrics archive [158]		2678	10368
Teuvo Kohonen citations [158]		3704	12673
Joshua Lederberg citations [158]		8212	41430
Ahmed Zewail citations [158]		6640	54173
High E. Particle Phys. archive [86]		27400	352021
Mobile phone records [162]	Communication	345	355
Emails at a university [68]		1133	5451
Emails at *Enron* [103]		33696	180811
Novel *David Copperfield* [154]	Information	112	425
Roget's Thesaurus dictionary [163]		994	3640
Java documentation (`javax`) [164]		1031	4408
ODLIS dictionary [165]		2898	16376
USF association norms [166]		10617	63782
FOLDOC dictionary [99]		13356	91471
WordNet dictionary [158]		75606	119564
Small software project [158]	Software	83	125
JUNG graph framework [167]		398	943
Java language (`javax`) [167]		1570	7194
Java language (general) [158]		1538	7817
Oregon aut. systems [98]	Internet	22963	48436
Gnutella file sharing [91]		36646	88303

Table 4.2

Real-world networks. (n and m correspond to the number of nodes and links, respectively.)

Network	Type	n	m
European roads [85]	Technological	1039	1305
Finite automaton [158]		1096	1677
US air lines [158]		332	2126
US power grid [158]		4941	6594
Escherichia Coli regulatory [158]	Biological	328	456
Caenorhabditis Elegans neural [22]		297	2148
Yeast protein interactions [168]		2224	6609
Data modeling [158]	Other	638	1020
Amazon products [96]	Co-purchase	524366	1491774
nd.edu domain [107]	Web graph	325729	1497135
Pennsylvania roads [103]	Technological	1087562	1541514
Wikipedia talk service [100]	Communication	2388953	4656682
Skitter overlay map [87]	Internet	1694616	11094209

particular, d is expressed as a power function of n through formula $d = c \cdot n^{-\gamma}$, where γ is a scaling exponent and c is a constant. We measure goodness of fit to the data using coefficient of determination R^2—how well the network size predicts density—and dependence between both variables corresponding to Spearman's correlation coefficient ρ—the extent to which network density decreases as network size increases. Moreover, we also evaluate the number of self-similar scales S defining how many renormalized networks are revealed under different techniques.

Mean estimates for each method appear in Tab. 4.3. Coefficients R^2 expose that the power-law relationship between the size and density appears to be a good fit to the data under box-covering methods and balanced propagation based renormalization. (We can reject the null hypothesis—no actual relationship between variables—at one percent significance level, thus results are statistically significant.) Irrespective of renormalization technique, R^2 and ρ for original networks are improved considering also their renormalized varieties. Otherwise, box-covering methods perform better than community detection algorithms, whereas balanced propagation exhibits the most homogeneous relationship between size and density. Spectral algorithm and modularity optimization prove the worst, particularly at observing fits for renor-

Table 4.3

Estimates of the fit for power-law scaling of network density and size in 50 real-world networks revealed under different renormalization techniques. Values are estimates of the mean over 10 renormalizations of each network and correspond to correlation coefficient ρ, coefficient of determination R^2, expressed network density d and the number of revealed self-similar scales S. (For each technique, ρ and R^2 are exposed separately for original and renormalized networks, and for renormalized varieties only—first and second row, respectively. Bold values of R^2 indicate relatively high goodness of fit to a power-law, whereas values in italics show poor performance of the respective renormalization technique.)

Technique	ρ	R^2	d	S
Randomized box-tiling	−0.975 −0.973	0.944 0.936	$1.7 \cdot n^{-0.807}$	5.3
Randomized cluster-growing	−0.977 −0.977	0.948 0.944	$1.6 \cdot n^{-0.818}$	4.6
Balanced propagation	−0.985 −0.980	0.962 0.963	$1.9 \cdot n^{-0.836}$	4.3
Modularity optimization	−0.966 −0.889	0.956 *0.820*	$3.0 \cdot n^{-0.882}$	3.9
Spectral analysis	−0.951 −0.878	0.922 *0.718*	$4.1 \cdot n^{-0.893}$	4.5
Original networks	−0.924	0.870	$3.8 \cdot n^{-0.921}$	

malized networks only. In the case of modularity optimization, this could be largely due to its resolution limit [152] and other weaknesses [153]. On the other hand, spectral analysis is in fact an optimization of eigenvectors of the modularity matrix. Therefore, it is attributed to the above mentioned modularity limitations, whereas it also reveals modules in random networks [169].

The plots on Fig. 4.2 – 4.3 illustrate size and density relationships with the scaling exponents γ around −0.85. Original networks exhibit greater scaling factor (see also section 4.3.3), which indicates γ is approaching −1 for adequately large n. This corresponds to commonly observed finding that most large-scale real-world networks tend to be sparse—the number of links appears not to be close to $O(n^2)$ but rather of order $O(n)$. Consecutively, we can simplify density definition with the relationship $d \approx n^{-1}$. Thus, power-law relationship between the network size and density is expected for original networks (without considering reduced varieties). However, among renormalized networks the relationships follow even stronger power-laws. This means that networks obtained on different scales of renormalization process also satisfy power-law relationship between size and density, and implies an existence

of density scaling also within real-world networks.

Furthermore, results show similar behavior of exponents γ and constants c for better performing techniques, including box-tiling, cluster-growing, and balanced propagation. This finding implies that box-covering methods find smaller and sparser boxes, similar to communities detected with balanced propagation. Other two algorithms reveal bigger, denser, and also more heterogeneous communities considering density scaling. The values of self-similar scales S are in accordance with these observations. Modularity optimization extracts network with one community in the least number of scales on average (bigger communities). On the other hand, box-tiling obtains a larger number of reduced networks (smaller boxes), which is expected due to the distance l_B setting.

To summarize, the analysis of real-world networks reveals power-law scaling of the network density with respect to network size. Among the employed renormalization techniques, balanced propagation seems to lead to the most optimal reduction of networks according to the density scaling. Results acquired by three best performing techniques indicate an existence of a certain common organizing principle of networks, which dictates linking rules and interactions among nodes. Our findings thus advance a recent discovery of a universal scaling of density among real-world networks [24], since we reveal density scaling also among different self-similar scales of complex real-world networks. In addition, the results are consistent with the densification laws of Leskovec et al. [87, 91]—$m \propto n^{\alpha}$, where α ranges between 1 and 2 and relates with our exponent γ, which lies between 0 and -1 respectively. Thus, our study expands densification laws to other dimensions of network structure.

Besides density, we also studied the scaling of other network properties with respect to network size. In particular, we analyzed number of links, average and maximum degree, number of articulation points, average path and diameter [22], betweenness and closeness centrality [170] and clustering coefficient [66]. The results reveal significant scaling also between network size and average node or link betweenness—the number of shortest paths going through a node and link respectively. Regarding to a definition of network density and observed power-law relationship between size and density, similar relationship for number of links occurs expectedly. However, due to simplicity, detailed investigation of betweenness centrality scaling is omitted, although a prominent direction for future research.

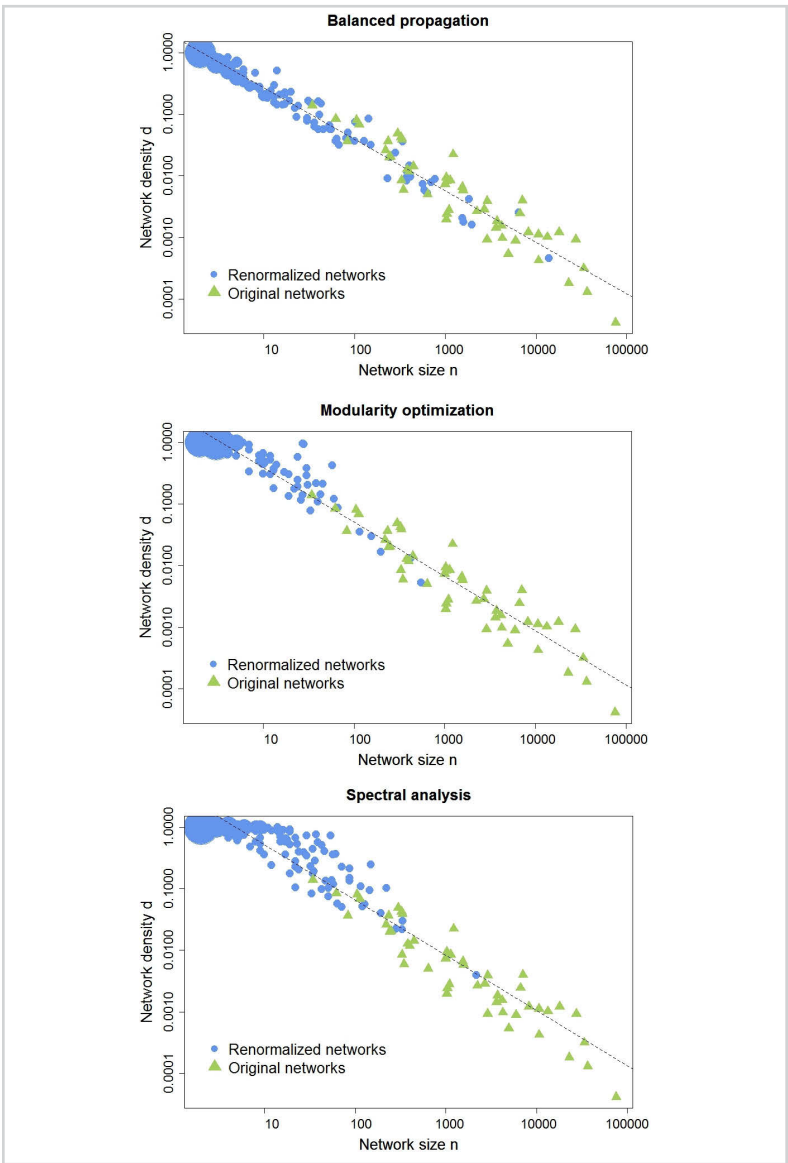


Figure 4.2

Power-law scaling of network density and size in 50 real-world networks of diverse origin revealed with different renormalization techniques. Plots show scaling of density for a single renormalization of each network under respective technique. (Green triangles correspond to original networks, whereas blue circles represent their renormalized varieties. Symbol sizes are proportional to the number of networks with the same size and density.)

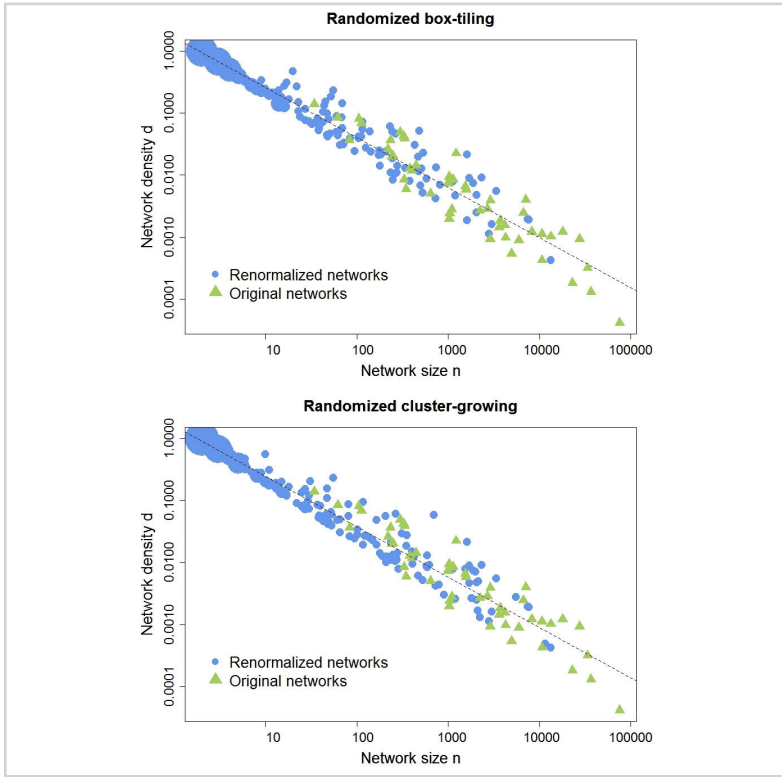


Figure 4.3

Power-law scaling of network density and size in 50 real-world networks of diverse origin revealed with different renormalization techniques. Plots show scaling of density for a single renormalization of each network under respective technique. (Green triangles correspond to original networks, whereas blue circles represent their renormalized varieties. Symbol sizes are proportional to the number of networks with the same size and density.)

4.3.2 Random networks

To further validate our results we apply box-tiling and modularity optimization renormalizations to Erdős-Rényi random graphs with different sizes n and probabilities of linking nodes p. We generate networks with 500, 1.000, 2.500, 5.000, and 10.000 nodes and probabilities corresponding to density obtained with modularity optimization based renormalization (section 4.3.1), density reported in [24], and probability that should assure sufficient size of the largest network component [156].

Firstly, we test balanced propagation renormalization, since the method performs best on real-world networks. The results prove to be very good,

showing fits closely to ideal (R^2 and ρ close to 1 and -1, respectively). However, detailed investigation shows renormalization for most of the generated networks reveals only a single scale or concludes without reduction, since random networks supposedly have no community structure. For this reason we exclude balanced propagation from the analysis. Thus, we study box-tiling as an illustration of classical box-covering principle and modularity optimization as an example of community based renormalization. Note that, in contrast to the above, the latter reveals non-trivial modules also in random networks (section 4.3.1).

The results appear in Tab. 4.4. A strong relationship ($R^2 = 1$, $\rho = -1$) arises between size and density of original networks. That occurs due to the settings of probability p. These strong fits cause also high values for original and randomized networks together. The results for randomized varieties of networks show low fits to the data and implie rather diverse density of reduced networks with respect to their size. This is anticipated owing to random network structure. However, the values of R^2 and ρ for randomized networks under $p = 2/(n-1)$ setting are relatively high. Examining plot for box-tiling closely shows diverse density among reduced networks, however, diversity straightens due to the large number of reduction scales. On the other hand, networks reduced under modularity optimization on each scale reveal almost the same density, and thus lead to higher fit. Slightly greater values for renormalized networks under box-tiling seem to occur due to the definition of boxes, which consider only proximity among nodes.

Other variables, including scaling exponent γ, constant c, and revealed self-similar scales S, comprehend greater range than values for real-world networks. This verifies there exists no optimal density characteristic for random networks and denotes that random networks do not exhibit common power-law density scaling.

According to the above, we conclude that results for random networks appear to be weak as anticipated, since random networks should not reveal structures like communities in real-world networks. On the contrary, findings for random networks indicate that self-similar density scaling of real-world networks is not obtained by chance, and the scaling exists due to some inner principles which determine network structure.

Table 4.4

Estimates of the fit for power-law scaling of network density and size in Erdős-Rényi random graphs obtained with two renormalization techniques. For each probability of a link between two nodes p, we construct an ensemble of networks of various sizes. Values are estimates of the mean over 10 realizations of each random graph. (See also Tab. 4.3.)

p	Technique	ρ	R^2	d	S
$3.0 \cdot n^{-0.882}$	Randomized box-tiling	−0.925 −0.852	*0.820* *0.787*	$2.8 \cdot n^{-0.753}$	4.9
	Modularity optimization	−0.957 −0.583	0.994 *0.446*	$12.3 \cdot n^{-0.963}$	3.0
$7.9 \cdot n^{-0.986}$	Randomized box-tiling	−0.939 −0.882	*0.818* *0.781*	$3.7 \cdot n^{-0.797}$	4.5
	Modularity optimization	−0.964 −0.494	0.998 *0.537*	$10.5 \cdot n^{-1.022}$	3.0
$2/(n-1)$	Randomized box-tiling	−0.990 −0.986	0.967 0.962	$2.6 \cdot n^{-0.953}$	6.7
	Modularity optimization	−0.930 −0.744	0.916 *0.817*	$6.4 \cdot n^{-1.065}$	4.0

4.3.3 Large real-world networks

For a complete analysis, we also analyze the size and density relationship of the largest five real-world networks presented in Tab. 4.2. In particular, co-purchase network of different products from Amazon in 2006, complete map of *nd.edu* domain, road network of Pennsylvania, communication network of user discussions on Wikipedia before January 2008, and internet topology graph from traceroutes in 2005. Due to simplicity, we present study only for the best preforming balanced propagation based renormalization, where the maximum number of iterations is limited to 100.

The results are presented in Tab. 4.5. Observing only original networks, fits are expectedly low due to small number of networks considered. For the same reason the constant c and exponent γ also differ from the ones in section 4.3.1. However, other results show very good fit particularly for original and randomized networks together and reveal a power-law relationship of network size and density (see Fig. 4.4). (Again, the results are statistically significant at one percent significance level.) As expected due to the size of the networks, the scaling exponent is close to −1. Number of self-similar scales

Table 4.5

Estimates of the fit for power-law scaling of network density and size in five large real-world networks revealed with balanced propagation. Values are estimates of the mean over 10 renormalizations of each network. (See also Tab. 4.3.)

Technique	ρ	R^2	d	S
Balanced propagation	−0.990 −0.980	0.977 0.961	$2.9 \cdot n^{-0.926}$	4.9
Original networks	−0.900	*0.719*	$66.2 \cdot n^{-1.175}$	

is higher as in analysis in section 4.3.1, since networks are larger and thus reduced in more steps. On the other hand, S does not significantly increase with network size, which implies that renormalization is effective and efficient approach for simplifying large networks.

Fig. 4.4 illustrates renormalized varieties of three large networks. We consider networks of diverse origin to value how different structure of networks effects the relationship between size and density. For instance, Pennsylvania roads network shows very homogeneous structure, while, on the contrary, other two networks present core-periphery structure typical for social and information networks. However, these diverse network structures do not reflect in the results (Fig. 4.4(a)). Thus, the finding confirms common density scaling in real-world networks irrespective of network type and origin.

Our study improves the comprehension of self-similar structure in real-world networks and implies several possible applications. Firstly, adequate network coarse-graining implies simplification and abstraction of large real-world networks without losing information about original network density. Reduction also enables visualization and improves the comprehension of larger complex networks. Additionally, self-similar density scaling can help at detecting sufficient density according to the size of the sub-graphs in graph sampling applications (e.g., [40]), improve the accuracy of link prediction (for review see [171]) and the quality of synthetic graph generation (e.g., [172]).

4.4 Conclusions

The paper explores the relationship between size and density of complex real-world networks under different box-covering and community-based renor-

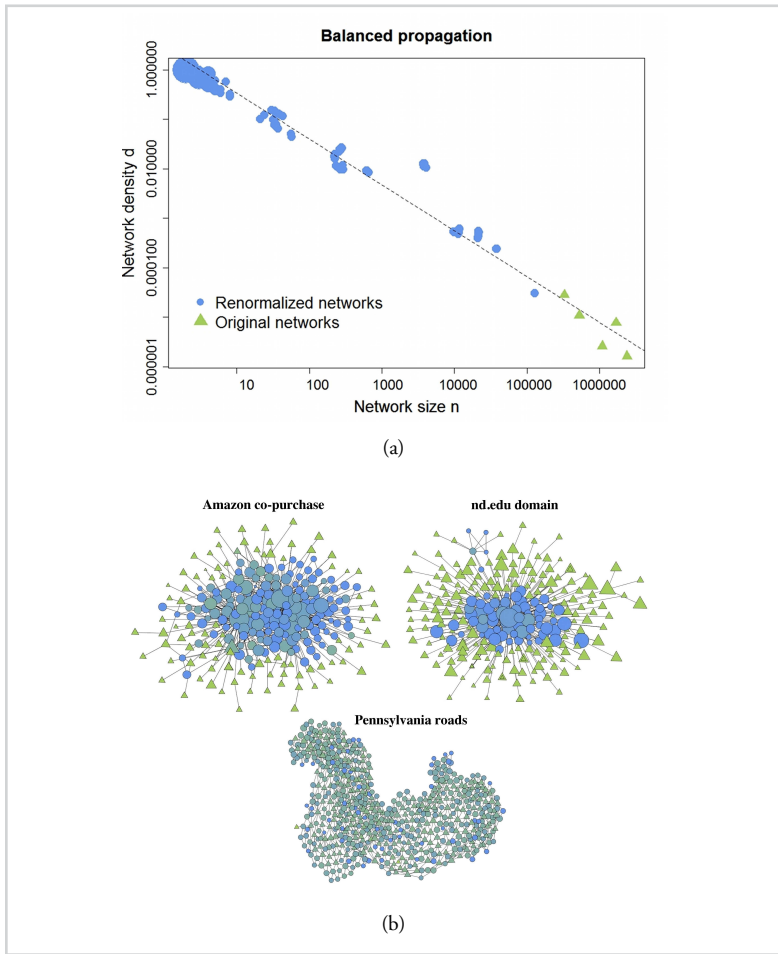


Figure 4.4

(a) Power-law scaling of network density and size in five real-world networks with millions of links revealed with balanced propagation. Plot shows scaling of density over 10 renormalizations of each network. (Green triangles correspond to original networks, whereas blue circles represent their renormalized varieties. Symbol sizes are proportional to the number of networks with the same size and density.) (b) Density of network structure in renormalized varieties of three large real-world systems of different origin. (Node symbols correspond to degree-corrected clustering coefficient [116] that ranges between 0 and 1—green triangles and blue circles, respectively—while symbol sizes are proportional to the number of nodes in the original network.)

malization techniques. The analysis was conducted on over 50 real-world networks of various sizes as well as different Erdős-Rényi random graphs. The main contribution of the study is to imply an existence of a scale-free density not only among different real-world networks, but also among their self-similar scales. Common scaling of density thus appears to be a unique property of complex real-world networks irrespective of their type, size and origin. Also, the results reveal balanced propagation based renormalization as the best performing method among the observed algorithms. The study on Erdős-Rényi random graphs, which supposedly exhibit no community structure, validates the above results and confirms that observed scaling of density is distinctive for real-world networks. Hence, our findings expand recent discoveries to other dimensions of network structure and further improve the comprehension of self-similarity in complex real-world networks. The latter has possible applications in graph sampling, link prediction, synthetic graph generation, network abstraction and visualization.

In our future work we intend to focus on other possible characteristics of density scaling, that could be identified in networks of common type and origin. Furthermore, we will analyze the betweenness centrality scaling with respect to network size in detail. Moreover, the work will also be extended on finding suitable ways for abstracting large real-world networks, while at the same time preserving their fundamental properties.

Acknowledgment

Authors would like to thank Matija Polajnar for providing Slovenian scientists collaboration data and Bojan Klemenc for useful comments. The work has been supported by the Slovene Research Agency *ARRS* within the research program P2-0359.

5

Zgoščevanje skupin vozlišč pri zmanjševanju družbenih in informacijskih omrežij

Any network studied in the literature is inevitably just a sampled representative of its real-world analogue. Additionally, network sampling is lately often applied to large networks to allow for their faster and more efficient analysis. Nevertheless, the changes in network structure introduced by sampling are still far from understood. In this paper, we study the presence of characteristic groups of nodes in sampled social and information networks. We consider different network sampling techniques including random node and link selection, network exploration and expansion. We first observe that the structure of social networks reveals densely linked groups like communities, while the structure of information networks is better described by modules of structurally equivalent nodes. However, despite these notable differences, the structure of sampled networks exhibits stronger characterization by community-like groups than the original networks, irrespective of their type and consistently across various sampling techniques. Hence, rich community structure commonly observed in social and information networks is to some extent merely an artifact of sampling.

5.1 Introduction

Any network found in the literature is inevitably just a sampled representative of its real-world analogue under study. For instance, many networks change quickly over time and in most cases merely incomplete data is available on the underlying system. Additionally, network sampling techniques are lately often applied to large networks to allow for their faster and more efficient analysis. Since the findings of the analyses and simulations on such sampled networks are implied for the original ones, it is of key importance to understand the structural differences between the original networks and their sampled variants.

A large number of studies on network sampling focused on the changes in network properties introduced by sampling. Lee et al. [57] showed that random node and link selection overestimate the scale-free exponent [115] of the degree and betweenness centrality [35] distributions, while they preserve the degree mixing [65]. On the other hand, random node selection preserves the degree distribution of different random graphs [52] and performs better for larger sampled networks [53]. Furthermore, Leskovec et al. [40] showed that the exploration sampling using random walks or forest-fire strategy [87] outperforms the random selection techniques in preserving the clustering coeffi-

cient [22], different spectral properties [40], and the in-degree and out-degree distributions. More recently, Ahmed et al. [59] proposed random link selection with additional induction step, which notably improves on the current state-of-the-art. Their results confirm that the proposed technique well captures the degree distributions, shortest paths [22] and also the clustering coefficient of the original networks. Lately, different studies also focus on finding and correcting biases in sampling process, for example observing the changes of user attributes under the sampling of social networks [173], analyzing the bias of traceroute sampling [174] and understanding the changes of degree distribution and hubs inclusion under various sampling techniques [175]. However, despite all those efforts, the changes in network structure introduced by sampling and the effects of network structure on the performance of sampling are still far from understood.

Real-world networks commonly reveal communities (also link-density community [176]), described as densely connected clusters of nodes that are loosely connected between [27]. Communities possibly play important roles in different real-world systems, for example in social networks communities represent friendship circles or people with similar interest [25], while in citation networks communities can help us to reveal relationships between scientific disciplines [26]. Furthermore, community structure has a strong impact on dynamic processes taking place on networks [31] and thus provides an important insight into structural organization and functional behavior of real-world systems. Consequently, a number of community detection algorithms have been proposed over the last years [149, 164, 177, 178] (for a review see [30]). Most of these studies focus on classical communities characterized by higher density of edges [179]. However, some recent works demonstrate that real-world networks reveal also other characteristic groups of nodes [32, 33] like groups of structurally equivalent nodes denoted modules [32, 34] (also link-pattern community [176] and other [180]), or different mixtures of communities and modules [73].

Despite community structure appears to be an intrinsic property of many real-world networks, only a few studies considered the effects between the community structure and network sampling. Salehi et al. [71] proposed PageRank sampling, which improves the performance of sampling of networks with strong community structure. Furthermore, expansion sampling [72] directly constructs a sample representative of the community structure, while it can also be used to infer communities of the unsampled nodes. Other studies,

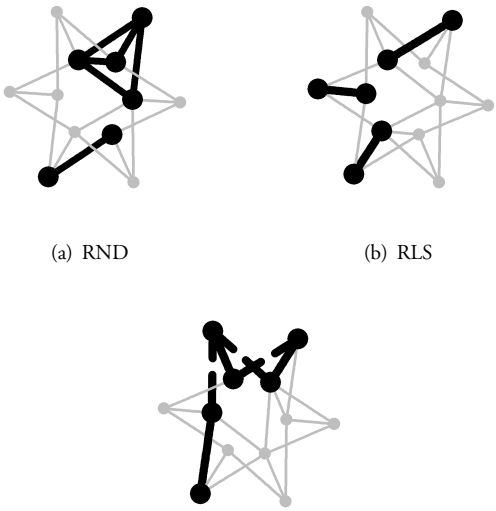
for example analyzed the evolution of community structure in collaboration networks and showed that the number of communities and their size increase over time [181], while the network sampling has a potential application in testing for signs of preferential attachment in the growth of networks [182]. However, to the best of our knowledge, the question whether sampling destroys the structure of communities and other groups of nodes or are sampled nodes organized in a similar way than nodes in original network remains unanswered.

In this paper, we study the presence of characteristic groups of nodes in different social and information networks and analyze the changes in network group structure introduced by sampling. We consider six sampling techniques including random node and link selection, network exploration and expansion sampling. The results first reveal that nodes in social networks form densely linked community-like groups, while the structure of information networks is better described by modules. However, regardless of the type of the network and consistently across different sampling techniques, the structure of sampled networks exhibits much stronger characterization by community-like groups than the original networks. We therefore conclude that the rich community structure is not necessary a result of for example homophily in social networks.

The rest of the paper is structured as follows. In Section 5.2, we introduce different sampling techniques considered in the study, while the adopted node group extraction framework is presented in Section 5.3. The results of the empirical analysis are reported and formally discussed in Section 5.4, while Section 5.5 summarizes the paper and gives some prominent directions for future research.

5.2 *Network sampling*

Network sampling techniques can be roughly divided into two categories: random selection and network exploration techniques. In the first category, nodes or links are included in the sample uniformly at random or proportional to some particular characteristic like the degree of a node or its PageRank score [183]. In the second category, the sample is constructed by retrieving a neighborhood of a randomly selected seed node using random walks, breadth-first search or another strategy. For the purpose of this study, we consider three techniques from each of the categories.



(a) RND (b) RLS

(c) RLI

Figure 5.1

Random selection techniques applied to a small toy network, where the samples are shown with highlighted nodes and links. (a) In random node selection by degree, the nodes are selected with probability proportional to their degrees, while their mutual links are included in the sample. (b) In random link selection, the sample consists of links selected uniformly at random. (c) In random link selection with induction, the sample consists of randomly selected links (solid lines) and the links between their endpoints (dashed lines).

5.2.1 Random selection

From the random selection category, we first adopt random node selection by degree [40] (RND). Here, the nodes are selected randomly with probability proportional to their degrees, while all their mutual links are included in the sample (Fig. 5.1(a)). Note that RND improves the performance of the basic random node selection [40, 60], where the nodes are selected to the sample uniformly at random. RND fits better spectral network properties [40] and produces the sample with larger weakly connected component [60]. Moreover, it shows good performance in preserving the clustering coefficient and betweenness centrality distribution of the original networks [60]. Nevertheless, it can still construct a disconnected sample network, despite a fully connected original network.

Next, we adopt random link selection [40] (RLS), where the sample consists of links selected uniformly at random (Fig. 5.1(b)). RLS overestimates

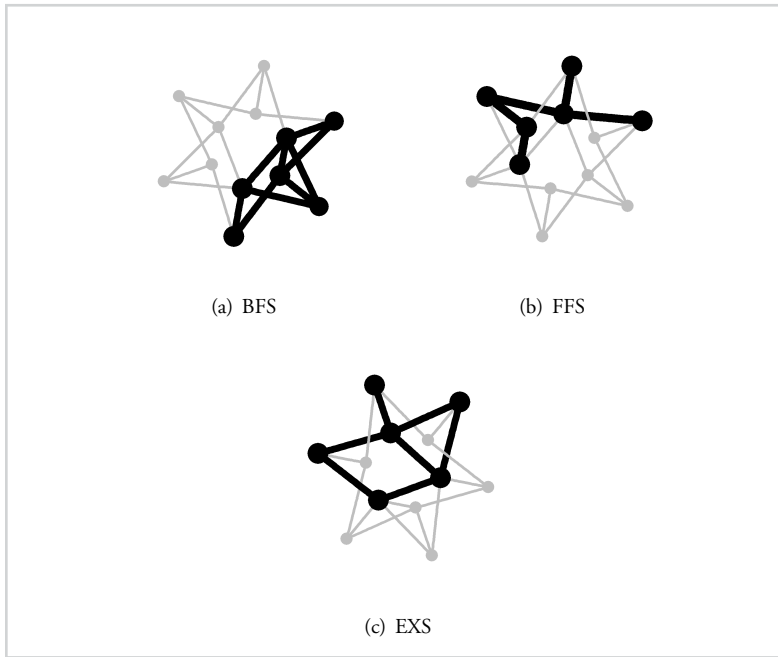


Figure 5.2

Network exploration techniques applied to a small toy network, where the samples are shown with highlighted nodes and links. (a) In breadth-first sampling, a seed node is first selected uniformly at random, while its broad neighborhood retrieved from breadth-first search is included in the sample. (b) In forest-fire sampling, the broad neighborhood of a randomly selected seed node is retrieved from partial breadth-first search, where only a fraction of neighbors is included in the sample. (c) In expansion sampling, the seed node is selected uniformly at random, while the remaining nodes are selected from the neighborhood of sampled nodes with probability proportional to their contribution to the expansion factor (see text).

degree and betweenness centrality exponent, underestimate the clustering coefficient and accurately matches the assortativity of the original network [57]. The samples created with RLS are sparse and the connectivity of the original network is not preserved, still RLS is likely to capture the path length of the original network [184].

Last, we adopt random link selection with induction [59] (RLI), which improves the performance of RLS. In RLI, the sample consists of randomly selected links as before, while also all additional links between their endpoints (Fig. 5.1(c)). RLI outperforms several other methods in capturing the degree, path length and clustering coefficient distribution. It selects nodes with higher

degree than RLS, thus the connectivity of the sample is increased [59].

Techniques from random selection category imitate classical statistical sampling approaches, where each individual is selected from population independently from others until desired size of the sample is reached.

5.2.2 *Network exploration*

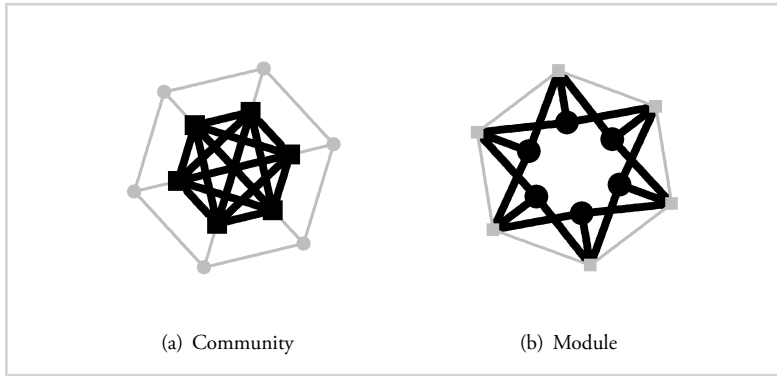
From the network exploration category, we first adopt breadth-first sampling [57] (BFS). Here, a seed node is selected uniformly at random, while its broad neighborhood retrieved from the basic breadth-first search is included in the sample (Fig. 5.2(a)). The sample network is thus a connected subgraph of the original network. BFS is biased towards selecting high-degree nodes in the sample [41]. It captures well the degree distribution of the networks, while it performs worst in inclusion of hubs in the sample quickly in the sampling process [175]. BFS imitates the snowball sampling approach for collecting social data used especially when the data is difficult to reach [185]. Selected seed participant is asked to report his friends, which are than invited to report their friends. The procedure is repeated until the desired number of people is sampled.

Next, we adopt a modification of BFS denoted forest-fire sampling [40] (FFS). In FFS, the broad neighborhood of a randomly selected seed node is retrieved from partial breadth-first search, where only some neighbors are included in the sample on each step (Fig. 5.2(b)). The number of neighbors is sampled from a geometric distribution with mean $p/(1-p)$, where p is set to 0.7 [40]. FFS matches well spectral properties [40], while it underestimates the degree distribution and fails to match the path length and clustering coefficient of the original networks [184]. However, FFS corresponds to a model by which one author collects the papers to cite and include them in the bibliography [87]. The author starts with one paper, explores its bibliography and selects the papers to cite. The procedure is recursively repeated in selected papers until desired collection of citations is reached.

Last, we adopt expansion sampling [72] (EXS), where the seed node is again selected uniformly at random, while the neighbors of the sampled nodes are included in the sample with probability proportional to

$$1 - \beta^{|N(\{v\}) - (N(S) \cup S)|}, \tag{5.1}$$

where v is the concerned node, S the current sample and $N(S)$ the neighborhood of nodes in S (Fig. 5.2(c)). Expression $|N(\{v\}) - (N(S) \cup S)|$ denotes



(a) Community (b) Module

Figure 5.3

Toy examples of groups of nodes in networks, where groups S and their corresponding linking patterns T are shown with highlighted and squared nodes, respectively (see text). (a) Communities are densely connected groups of nodes with $S = T$. (b) Modules are possibly disconnected groups of structurally equivalent nodes with $S \cap T = \varnothing$. Groups spanning between communities and modules are denoted mixtures.

the expansion factor of node v for sample S and means the number of new neighbors contributed by v. The parameter β is set to 0.9 [72]. Note that EXS ensures that the sample consists of nodes from most communities in the original network and that the nodes that are grouped together in the original network, are also grouped together in the sample [27]. EXS imitates the modification of snowball sampling approach mentioned above, where for example we want to gather the data about individuals from different countries. Thus, on each step we include in the sample the individuals, which knows larger number of others from various countries.

5.3 Group extraction

The node group structure of different networks is explored by a group extraction framework [73, 74, 186] with a brief overview below.

Let the network be represented by an undirected graph $G(V, L)$, where V is the set of nodes and L the set of links. Next, let S be a group of nodes and T a subset of nodes representing its corresponding linking pattern (i.e., the pattern of connections of nodes from S to other nodes [32]), $S, T \subseteq V$. Denote $s = |S|$ and $t = |T|$. The linking pattern T is selected to maximize the number of links between S and T, and minimize the number of links between

S and T^C, while disregarding the links with both endpoints in S^C. For details on the group objective function see [73, 187].

The above formalism comprises different types of groups commonly analyzed in the literature (Fig. 5.3). It consider communities [27] (i.e., link-density community [176]), defined as a (connected) group of nodes with more links toward the nodes in the group than to the rest of the network [179]. Communities are characterized by $S = T$. Furthermore, the formalism consider possibly disconnected groups of structurally equivalent nodes denoted modules [32, 34] (i.e., link-pattern community [176]), defined as a (possibly) disconnected group of nodes with more links towards common neighbors than to the rest of the network [179]. Modules have $S \cap T = \emptyset$. Communities and modules represent two extreme cases with all other groups being the mixtures of the two [73], $S \cap T \subset S$ and/or $S \cap T \subset T$. The reader may also find it interesting that the core-periphery structure is a mixture with $S \subset T$, while the hub & spokes structure is a module with $t = 1$.

The type of group S can in fact be determined by the Jaccard index [188] of S and its corresponding linking pattern T. The group parameter τ [73], $\tau \in [0, 1]$, is defined as

$$\tau(S, T) = \frac{|S \cap T|}{|S \cup T|}. \tag{5.2}$$

Communities have $\tau = 1$, while modules are indicated by $\tau = 0$. Mixtures correspond to groups with $0 < \tau < 1$. For the rest of the paper, we refer to groups with $\tau \approx 1$ as community-like and groups with $\tau \approx 0$ as module-like.

Groups in networks are revealed by a sequential extraction procedure proposed in [73, 74, 186]. One first finds the group S and its linking pattern T with random-restart hill climbing [189] that maximizes the objective function. Next, the revealed group S is extracted from the network by removing the links between groups S and T, and any node that becomes isolated. The procedure is then repeated on the remaining network until the objective function is larger than the 99th percentile of the values obtained under the same framework in a corresponding Erdős-Rényi random graph [156]. All groups reported in the paper are thus statistically significant at 1% level. Note that the above procedure allows for overlapping [190], hierarchical [191], nested and other classes of groups.

Table 5.1

Social and information networks considered in the study.

Network	Description	# Nodes	# Links
Collab	High Energy Physics collaborations [87]	9877	25998
PGP	Pretty Good Privacy web-of-trust [90]	10680	24340
P2P	Gnutella peer-to-peer file sharing [87]	8717	31525
Citation	High Energy Physics citations [87]	27770	352807

5.4 Analysis and discussion

Section 5.4.1 introduces real-world networks considered in the study. Section 5.4.2 reports the node group structure of the original networks extracted with the framework described in Section 5.3. The groups extracted from the sampled networks are analyzed in Section 5.4.3. For a complete analysis, we also observe the node group structure of a large network with more than a million links in Section 5.4.4.

5.4.1 Network data

The empirical analysis in the following sections was performed on four real-world social and information networks. Their main characteristics are shown in Table 5.1.

The *Collab* [87] is a social network of scientific collaborations among researchers, who submitted their papers to High Energy Physics – Theory category on the arXiv in the period from January 1993 to April 2003. The nodes represent the authors, while undirected links denote that two authors co-authored at least one paper together.

The *PGP* [90] is a social network, which corresponds to the interaction network of users of the Pretty Good Privacy algorithm collected in July 2001. The nodes represent users, while undirected links indicate relationships between those, who sign each other's public key.

The *P2P* [87] is an information network, which contains a sequence of snapshots of the Gnutella peer-to-peer file sharing network collected in August 2002. The nodes represent hosts in the Gnutella network, which are linked by undirected links if there exist connections between them.

The *Citation* [87] is an information network, again gathered from the

Table 5.2

Groups of nodes extracted from social and information networks. We report the number of groups #, the mean group size s, the mean pattern size t, the mean group parameter τ, the median group parameter denoted m_τ and the distribution over different types of groups (see text). Notice that social networks consist of smaller groups with larger τ than information networks.

Network	Group				
	#	$\langle s \rangle$	$\langle t \rangle$	$\langle \tau \rangle$	m_τ
Collab	129	66.9	67.2	0.568	0.554
PGP	87	62.2	61.9	0.568	0.536
P2P	70	154.8	177.0	0.057	0.000
Citation	284	271.7	280.6	0.186	0.120

Network	Community	Mixture	Module
	Distribution %		
Collab	1.6%	96.8%	1.6%
PGP	4.6%	94.3%	1.1%
P2P	0.0%	44.3%	55.7%
Citation	0.0%	96.8%	3.2%

High Energy Physics – Theory category from the arXiv in the period from January 1993 to April 2003 and includes the citations among all papers in the dataset. The network consists of nodes, which represent papers, while links denote that one paper cite another.

5.4.2 *Group structure of original networks*

We first analyze the properties of groups extracted from the original networks summarized in Table 5.2.

The number of groups differs among networks, still the mean group size s (denoted $\langle s \rangle$) is comparable across network types. Groups S in social networks consist of around 64 nodes, while $\langle s \rangle$ in information networks exceeds 150 nodes. The mean linking pattern size t (denoted $\langle t \rangle$) of social networks is comparable to $\langle s \rangle$. The latter relation $\langle t \rangle \approx \langle s \rangle$ is expected due to the pronounced community structure commonly found in social networks [192]. On the other hand, $\langle t \rangle > \langle s \rangle$ is expected for information networks, due to the abundance of module-like groups.

The characteristic group structure of networks is reflected in the group parameter τ. For social networks, its values are around 0.556, which indicates

the presence of communities, modules and mixtures of these. In contrast to social networks, the information networks have τ closer to 0 and consist mostly of module-like groups.

To summarize, social networks represent people and interactions between them, like a few authors writing a paper together, therefore we can expect a larger number of community-like groups in these networks. On the other hand, in information networks the homophily is less typical and thus the structure of these networks seem better described by module-like groups.

5.4.3 Group structure of sampled networks

Sampling techniques outlined in Section 5.2 enable setting the size of the sampled networks in advance. We consider sample sizes of 15% of nodes from the original networks, that provides for an accurate fit of several network properties [40, 60].

Table 5.3 and 5.4 present the properties of the node group structure of sampled social and information networks, respectively. Notice that RLS and FFS show different performance than other techniques. The samples obtained with RLS and FFS contain less groups, which consist of no more than 36 nodes. Additionally, almost all groups in these samples are modules, which reflects in the mean group parameter τ (denoted $\langle\tau\rangle$) approaching 0 for all networks.

To verify the above findings, we compute externally studentized residuals of the sampled networks that measure the consistency of each sampling technique with the rest. The residuals are calculated for each technique as the difference between the observed value of considered property and its mean divided by the standard deviation. The mean value and standard deviation are computed for all sampling techniques, excluding the observed one (for details see [193]). Statistically significant inconsistencies between techniques are revealed by two-tailed Student $t-$test [194] at $P-$value of 0.1, rejecting the null hypothesis that the values of the considered property are consistent across the sampling techniques.

Statistical comparison of sampling techniques for the number of groups and the mean group parameter τ is shown on Fig. 5.4. We confirm that the samples obtained with RLS and FFS reveal significantly less groups with significantly smaller $\langle\tau\rangle$ than other sampling techniques. Moreover, if we compare the number of links in the sampled networks, RLS and FFS create samples

Table 5.3

Groups of nodes extracted from sampled social networks over 100 realizations of different sampling techniques (see text). We report the number of groups # and standard deviation, the mean group size s, the mean pattern size t, the mean group parameter τ and standard deviation, the median group parameter denoted m_τ and the distribution over different types of groups. Notice that sampled networks expectedly consist of smaller groups, but with larger τ than original social networks (see $\langle\tau\rangle$ and m_τ).

Network	Sampling	Group #	$\langle s\rangle$	$\langle t\rangle$	$\langle\tau\rangle$	m_τ
Collab	/	129.0	66.9	67.2	0.568	0.554
	RND	65.4 ± 3.7	13.5	13.7	0.851 ± 0.030	0.989
	RLS	1.2 ± 0.5	1.5	4.8	0.047 ± 0.149	0.048
	RLI	74.7 ± 4.6	13.7	13.9	0.846 ± 0.030	0.979
	BFS	104.0 ± 6.5	18.2	18.5	0.787 ± 0.032	0.861
	FFS	4.0 ± 1.6	16.8	29.8	0.000 ± 0.000	0.000
	EXS	87.0 ± 5.8	18.4	18.9	0.741 ± 0.026	0.791
PGP	/	87.0	62.2	61.9	0.568	0.536
	RND	68.2 ± 4.5	15.8	16.0	0.891 ± 0.024	1.000
	RLS	2.8 ± 1.0	5.7	7.6	0.304 ± 0.233	0.263
	RLI	74.3 ± 4.3	15.8	16.1	0.883 ± 0.024	1.000
	BFS	95.4 ± 9.2	17.5	17.7	0.784 ± 0.025	0.909
	FFS	3.6 ± 1.3	13.5	32.6	0.000 ± 0.000	0.000
	EXS	80.9 ± 6.5	15.6	15.8	0.779 ± 0.028	0.873

Network	Sampling	Community	Mixture	Module
		Distribution %		
Collab	/	1.6%	96.8%	1.6%
	RND	54.7%	41.9%	3.4%
	RLS	0.0%	8.3%	91.7%
	RLI	52.7%	43.4%	3.9%
	BFS	30.3%	66.5%	3.2%
	FFS	0.0%	0.0%	100.0%
	EXS	21.4%	76.3%	2.3%
PGP	/	4.6%	94.3%	1.1%
	RND	67.8%	28.7%	3.5%
	RLS	21.4%	28.6%	50.0%
	RLI	65.1%	31.1%	3.8%
	BFS	39.2%	55.6%	5.2%
	FFS	0.0%	0.0%	100.0%
	EXS	34.5%	61.2%	4.3%

Table 5.4

Groups of nodes extracted from sampled information networks over 100 realizations of different sampling techniques (see text). We report the number of groups # and standard deviation, the mean group size s, the mean pattern size t, the mean group parameter τ and standard deviation, the median group parameter denoted m_τ and the distribution over different types of groups. Notice that sampled networks expectedly consist of smaller groups, but with larger τ than original information networks (see $\langle\tau\rangle$ and m_τ).

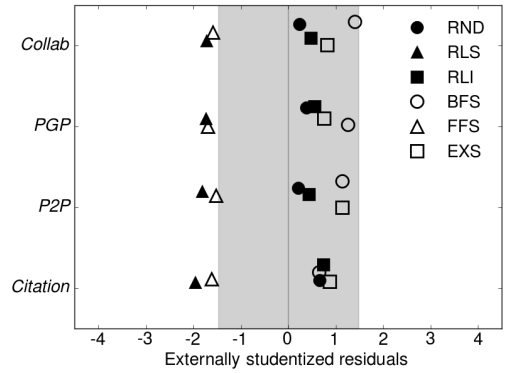
Network	Sampling	Group #	$\langle s\rangle$	$\langle t\rangle$	$\langle\tau\rangle$	m_τ
P2P	/	70.0	154.8	177.0	0.057	0.000
	RND	23.3 ± 3.9	24.2	24.4	0.163 ± 0.049	0.034
	RLS	1.6 ± 0.9	1.2	3.6	0.000 ± 0.008	0.000
	RLI	26.2 ± 4.4	27.5	28.1	0.161 ± 0.039	0.035
	BFS	34.1 ± 5.5	31.3	27.9	0.131 ± 0.042	0.034
	FFS	3.6 ± 1.4	17.8	28.3	0.000 ± 0.000	0.000
	EXS	34.0 ± 5.9	36.9	37.3	0.125 ± 0.030	0.035
Citation	/	284.0	271.7	280.6	0.186	0.120
	RND	121.4 ± 4.9	74.9	78.1	0.405 ± 0.016	0.329
	RLS	1.5 ± 1.2	1.4	15.3	0.014 ± 0.073	0.014
	RLI	124.8 ± 5.5	76.3	79.9	0.415 ± 0.014	0.344
	BFS	120.4 ± 7.1	99.2	100.9	0.359 ± 0.047	0.244
	FFS	10.6 ± 4.2	35.5	30.0	0.000 ± 0.000	0.000
	EXS	131.2 ± 6.0	91.4	95.4	0.388 ± 0.019	0.284

Network	Sampling	Community	Mixture	Module
		Distribution %		
P2P	/	0.0%	44.3%	55.7%
	RND	4.2%	45.8%	50.0%
	RLS	0.0%	0.0%	100.0%
	RLI	3.8%	48.8%	47.4%
	BFS	2.3%	50.7%	47.0%
	FFS	0.0%	0.0%	100.0%
	EXS	2.4%	53.8%	43.8%
Citation	/	0.0%	96.8%	3.2%
	RND	0.2%	80.9%	18.9%
	RLS	0.0%	0.0%	100.0%
	RLI	0.2%	82.6%	17.2%
	BFS	0.1%	77.5%	22.4%
	FFS	0.0%	0.0%	100.0%
	EXS	0.2%	82.0%	17.8%

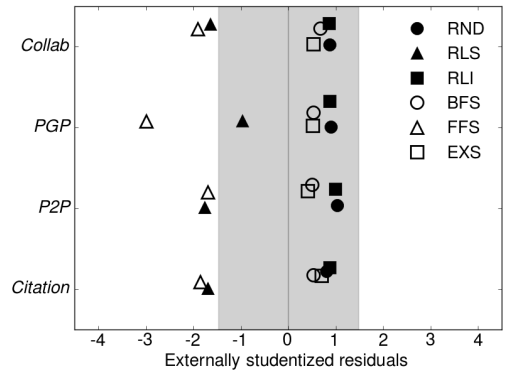
that contain on average 3% of links from the original networks. In contrast, the samples obtained with RND, RLI, BFS and EXS consist of around 16% of links from the original networks. As mentioned before, the sizes of all samples are 15% of the original networks, thus the sampled networks obtained with RLS and FFS are much sparser than others. In addition, the performance of RLS and FFS can also be explained by their definition. Since in RLS we include only randomly selected links in the sample, the variance is very high, while it commonly contains a large number of sparsely linked components, whose structure is best described as module-like. On the other hand, the samples obtained with FFS consist of one connected component with a low average degree of 2.33. Thus, the sparsely connected nodes also form groups, which are more similar to modules. Due to the above reasons, we exclude RLS and FFS from further analysis. We focus on RND, RLI, BFS, and EXS, whose performance is clearly more comparable.

The selected sampling techniques perform similarly across all networks as shown in Table 5.3 for social and Table 5.4 for information networks. The samples consist of various number of groups, still in most cases less than the original networks. The mean sizes s and t are around 40, in contrast to groups with 143 nodes on average in the original networks. Still, $\langle s \rangle \approx \langle t \rangle$ irrespective of network type and the sampling technique, which implies stronger characterization by community-like groups, as already argued in the case of social networks in Section 5.4.2.

Indeed, the majority of groups found in sampled social networks are community-like, which reflects in the parameter $\tau > 0.7$. In sampled information networks the number of mixtures decreases and communities appear, thus τ is larger than in the original networks. Fig. 5.5 – 5.6 shows a clear difference in the distribution of τ between the original and sampled networks. Furthermore, to confirm that differences exist between the structure of the original and sampled networks, we compute externally studentized residuals, where we include the value of considered property of the original network in computing the mean over different sampling techniques. We compare the number of groups and the parameter $\langle \tau \rangle$ for the original networks and their samples (Fig. 5.7). The results prove that the original networks contain a significantly larger number of groups with significantly smaller $\langle \tau \rangle$ than the sampled networks. Yet, larger parameter τ and consequently more community-like groups in sampled social networks and less module-like groups in sampled information networks indicate clear changes in the network structure intro-



(a) Number of groups #



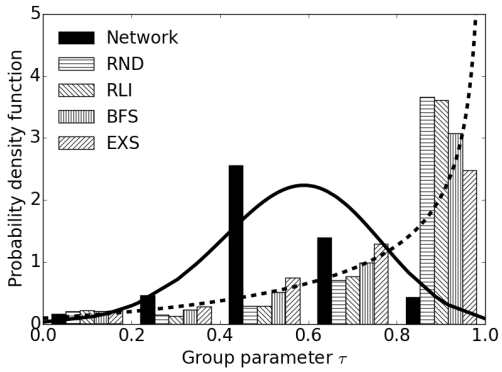
(b) Group parameter $\langle\tau\rangle$

Figure 5.4

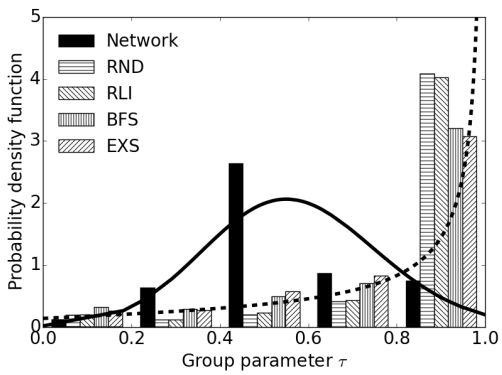
Statistical comparison of (a) number of groups and (b) mean group parameter τ for the sampled networks obtained with different sampling techniques (see text). We show externally studentized residuals that measure the consistency of each sampling technique with the rest and expose statistically significant inconsistencies between the techniques with two-tailed Student t-test at P-value of 0.1 (shaded regions correspond to 90% confidence intervals). Notice that sampled networks obtained with RLS and FFS reveal less groups (see (a)) with significantly smaller parameter τ (see (b)) than other sampling techniques.

duced by sampling. We conclude that these changes occur regardless of the network type or the adopted sampling technique.

Notice that the largest τ and thus the strongest characterization by community-like groups is revealed in the sampled networks obtained with both random selection techniques, RND and RLI. In RND nodes with higher degrees are more likely to be selected to the sample by definition, while RLI is biased in a similar way [59]. Thus, densely connected groups of nodes have



(a) *Collab* network

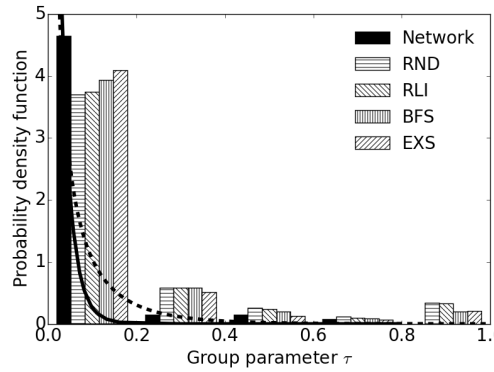


(b) *PGP* network

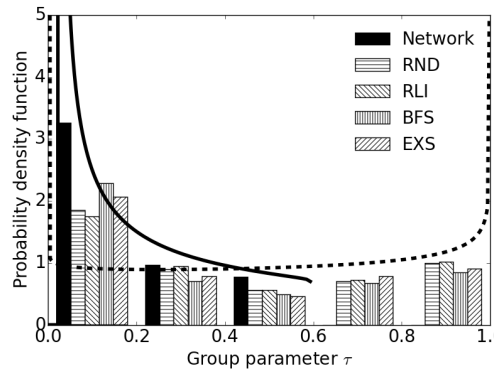
Figure 5.5

Distributions of group parameter τ for the original networks and their sampled representatives obtained with selected sampling techniques (see text). Histograms are derived by standard equidistant binning, while the estimates of a beta distribution for the original (solid lines) and sampled networks (dashed lines) are merely a guide for the eye. Notice that sampled networks are characterized by denser groups with notably larger τ than the original ones. Groups are more community-like in the case of social networks (see (a) and (b)), while less module-like in the case of information networks (see (a) and (b)).

a higher chance of being included in the sampled network, while sparse parts of the networks remain unsampled. On the other hand, BFS and EXS sample the broad neighborhood of a randomly selected seed node and thus the sampled network represents a connected component. In the case of BFS, all nodes and links of some particular part of the original network are sampled. The latter is believed to be representative of the entire network [41], yet BFS is biased towards sampling nodes with higher degree [195] and overestimates



(a) *P2P* network

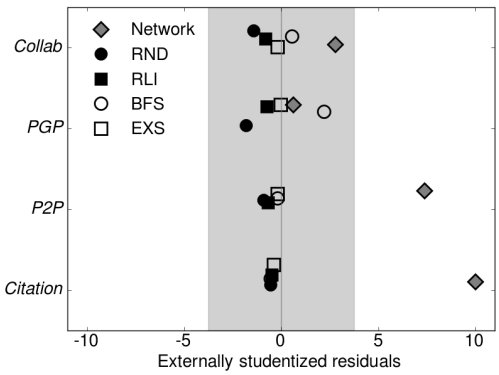


(b) *Citation* network

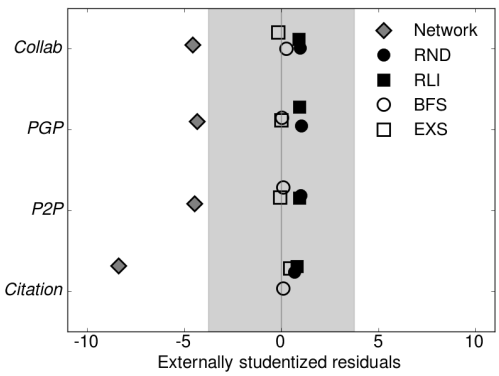
Figure 5.6

Distributions of group parameter τ for the original networks and their sampled representatives obtained with selected sampling techniques (see text). Histograms are derived by standard equidistant binning, while the estimates of a beta distribution for the original (solid lines) and sampled networks (dashed lines) are merely a guide for the eye. Notice that sampled networks are characterized by denser groups with notably larger τ than the original ones. Groups are more community-like in the case of social networks (see (a) and (b)), while less module-like in the case of information networks (see (a) and (b)).

the clustering coefficient, especially in information networks [57]. On the other hand, EXS ensures the smallest partition distance among several other sampling techniques, which means that nodes grouped together in communities of sampled network are also in the same community in the original network [72]. Therefore, the stronger characterization by community-like groups in sampled networks can also be explained by the definition and behavior of the sampling techniques.



(a) Number of groups #



(b) Group parameter $\langle\tau\rangle$

Figure 5.7

Statistical comparison of (a) number of groups and (b) mean group parameter τ for the original networks and their sampled representatives obtained with selected sampling techniques (see text). We show externally studentized residuals that measure the consistency of each network with the rest and expose statistically significant inconsistencies between the networks with two-tailed Student t-test at P-value of 0.1 (shaded regions correspond to 90% confidence intervals). Notice that original networks reveal more groups (see (a)) with significantly smaller parameter τ (see (b)) than the sampled networks.

5.4.4 *Group structure of a large network*

Due to the relatively high time complexity of the node group extraction framework, we consider only networks with a few thousand nodes. However, our previous study [60] proved that the size of the original network does not affect the accuracy of the sampling. Still, for a complete analysis, we also inspect the changes in node group structure introduced by sampling of a large *NotreDame*

Table 5.5

Groups of nodes extracted from the original *NotreDame* network and its sampled representatives over 100 realizations of selected sampling techniques (see text). We report the number of groups #, the mean group size s, the mean pattern size t, the mean group parameter τ and standard deviation, the median group parameter denoted m_τ and the distribution over different types of groups. Notice that sampled networks expectedly consist of smaller groups, but with larger τ than original network (see $\langle\tau\rangle$ and m_τ).

Sampling	Group				
	#	$\langle s\rangle$	$\langle t\rangle$	$\langle\tau\rangle$	m_τ
/	100	876.8	403.6	0.030	0.028
RND	100	302.5	271.7	0.369 ± 0.010	0.364
BFS	100	411.6	251.7	0.135 ± 0.030	0.119

Sampling	Community	Mixture	Module
	Distribution %		
/	0.0%	99.0%	1.0%
RND	0.0%	100.0%	0.0%
BFS	0.0%	99.5%	0.5%

network with more than a million links. Due to the simplicity and execution time, we present the analysis for two sampling techniques, RND from random selection and BFS from network exploration category. We also limit the number of groups extracted from the networks to 100 (i.e., we consider top 100 most significant groups with respect to the objective function).

The *NotreDame* data are collected from the web pages of the University of Notre Dame – *nd.edu* domain in 1999. The network contains 325,729 nodes representing individual web pages, while 1,497,134 links denote hyperlinks among them.

Table 5.5 shows the properties of groups, found in the original and sampled networks. The samples consist of smaller groups, still the mean size s remains larger than the mean size t. The majority of groups extracted from the original network are module-like, which reflects in the parameter τ slightly larger than 0. On the other hand, the changes introduced by sampling are clear, since the samples contain less modules, which is revealed by a larger parameter τ. These findings are consistent with the results on smaller networks from previous sections. The *NotreDame* as an information network expectedly consists of densely linked groups similar to modules, while the struc-

ture of sampled networks exhibits stronger characterization by community-like groups. That is again irrespective of the adopted sampling technique.

5.5 Conclusion

In this paper, we study the presence of characteristic groups of nodes like communities and modules in different social and information networks. We observe the groups of the original networks and analyze the changes in the group structure introduced by the network sampling.

The results first reveal noticeable differences in the group structure of original social and information networks. Nodes in social networks form smaller community-like groups, while information networks are better characterized by larger modules. After applying network sampling techniques, sampled networks expectedly contain fewer and smaller groups. However, the sampled networks exhibit stronger characterization by community-like groups than the original networks. We have shown that the changes in the node group structure introduced by sampling occur regardless of the network type and consistently across different sampling techniques. Since networks commonly considered in the literature are inevitably just a sampled representative of its real-world analogue, some results, such as rich community structure found in these networks, may be influenced by or are merely an artifact of sampling.

Our future work will mainly focus on larger real-world networks, including other types of networks like biological and technological. Moreover, we will further analyze the changes in the node group structure introduced by sampling and explore techniques that could overcome observed deficiencies.

Acknowledgment

This work has been supported in part by the Slovenian Research Agency *ARRS* within the Research Program No. P2-0359, by the Slovenian Ministry of Education, Science and Sport Grant No. 430-168/2013/91, and by the European Union, European Social Fund.

6

Zmanjševanje z indukcijo

Pristopi za zmanjševanje s preiskovanjem v splošnem ohranijo lastnosti omrežij bolje kot zmanjševanja z vzorčenjem [40]. Vpeljava dodatnega koraka indukcije v zmanjševanje z naključnim izbiranjem povezav izboljša delovanje osnovnega naključnega izbiranja [59]. Pri koraku indukcije v zmanjšano omrežje dodamo povezave iz osnovnega, s čimer se bolje ohrani porazdelitev stopenj vozlišč in nakopičenosti ter povprečna dolžina poti med vozlišči [59]. Po slednjem zgledu vpeljemo korak indukcije v preiskovanje z naključnim sprehodom in delno preiskovanje v širino. Z različnimi pristopi zmanjšamo več realnih omrežij ter opazujemo, kako se med zmanjševanjem spreminjajo njihove lastnosti pri različnih velikostih zmanjšanih omrežij. Delovanje pristopov z in brez indukcije primerjamo glede na ohranjanje porazdelitve stopenj vozlišč in nakopičenosti, povprečne stopnje in gostote omrežja.

6.1 Pristopi za zmanjševanje

Zmanjševanja z vzorčenjem in s preiskovanjem so sestavljena iz dveh korakov. V prvem koraku vzorčimo vozlišča ali povezave osnovnega omrežja. V drugem koraku iz izbranih vozlišč ali povezav sestavimo zmanjšano omrežje. Kadar pri tem uporabimo le vozlišča ali povezave, izbrane v prvem koraku, zmanjšanemu omrežju pravimo podgraf (angl. subgraph) osnovnega omrežja. Lahko pa pri sestavljanju zmanjšanega omrežja uporabimo v prvem koraku izbrana vozlišča oziroma krajišča izbranih povezav ter vse povezave, ki potekajo med njimi. V tem primeru zmanjšanemu omrežju pravimo induciran podgraf (angl. induced subgraph) osnovnega omrežja, korak dodajanja povezav med izbrana vozlišča pa imenujemo korak indukcije (angl. induction step).

V analizi primerjamo delovanje osmih pristopov za zmanjševanje. V nadaljevanju poglavja za posamezne pristope uporabljamo kratice, pojasnjene v tabeli 6.1. Štirje izmed pristopov temeljijo na vzorčenju. Pri naključnem izbiranju vozlišč [40] (angl. random node selection, RNS) na sliki 6.1(a) je zmanjšano omrežje sestavljeno iz naključno izbranih vozlišč in povezav med njimi. RNS dobro ohrani odvisnosti med stopnjami vozlišč [57] in razmerje gostote med osnovnim in zmanjšanim omrežjem [60]. Po drugi strani pa se pri zmanjševanju z RNS slabše ohrani nakopičenost [57], porazdelitev stopenj vozlišč [52] in povprečna dolžina poti med vozlišči [196]. Pri naključnem izbiranju vozlišč glede na stopnjo [40] (angl. random node selection based on degree, RND) na sliki 6.1(b) so vozlišča z večjo stopnjo v zmanjšano omrežje izbrana z večjo verjetnostjo. RND izboljša delovanje osnovnega naključnega

Tabela 6.1

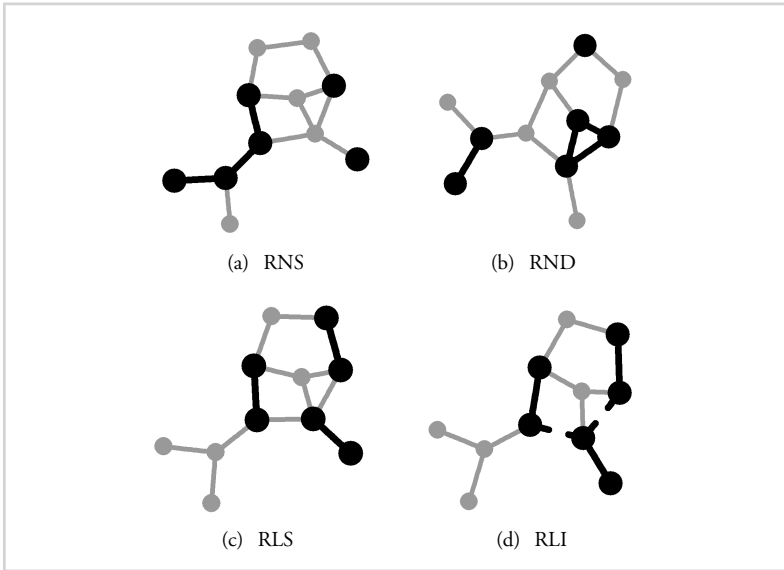
Kratice pristopov za zmanjševanje.

Kratica	Pristop
RNS	Naključno izbiranje vozišč
RND	Naključno izbiranje vozišč glede na stopnjo
RLS	Naključno izbiranje povezav
RLI	Naključno izbiranje povezav z indukcijo
BFS	Preiskovanje v širino
FFS	Delno preiskovanje v širino
FFI	Delno preiskovanje v širino z indukcijo
RWS	Preiskovanje z naključnim sprehodom
RWI	Preiskovanje z naključnim sprehodom z indukcijo
CGR	Združevanje vozlišč glede na razdaljo

izbiranja RNS [40]. Oba pristopa, RNS in RND, z veliko verjetnostjo ustvarita zmanjšano omrežje z več nepovezanimi komponentami.

Pri naključnem izbiranju povezav [40] (angl. random link selection, RLS) na sliki 6.1(c) je zmanjšano omrežje sestavljeno iz naključno izbranih povezav osnovnega omrežja. RLS dobro ohrani odvisnosti med stopnjami vozlišč [57] in povprečno dolžino poti med vozlišči [184]. Po drugi strani ustvari redka zmanjšana omrežja in z manjšo nakopičenostjo [57]. Delovanje RLS se izboljša z dodanim korakom indukcije [59] (angl. random link selection with subgraph induction, RLI), kjer zmanjšano omrežje sestavljajo naključno izbrane povezave, njihova krajišča ter vse povezave med krajišči, ki obstajajo v osnovnem omrežju (slika 6.1(d)). RLI deluje bolje kot nekateri drugi pristopi pri ohranjanju porazdelitve stopenj vozlišč, nakopičenosti in povprečne poti med vozlišči [59], hkrati pa z večjo verjetnostjo izbira vozlišča z večjo stopnjo, zaradi česar je zmanjšano omrežje sestavljeno iz več nepovezanih komponent [184].

V analizi uporabimo štiri pristope zmanjševanja s preiskovanjem. Preiskovanje z naključnim sprehodom [40] (angl. random walk sampling, RWS) simulira sprehod naključnega sprehajalca po omrežju, zmanjšano omrežje pa vsebuje vozlišča in povezave, ki jih naključni sprehajalec obišče (slika 6.2(a)). RWS dobro deluje na majhnih zmanjšanih omrežjih [40], dobro ohrani tranzitivnost [175] in porazdelitev nakopičenosti osnovnega omrežja, slabo pa ohrani porazdelitev stopenj vozlišč [41]. Pri delnem preiskovanju v širino [40] (angl. forest-fire sampling, FFS) z začetkom v naključno izbranem vozlišču preiskujemo omrežje v širino in na vsakem koraku v zmanjšano omrežje do-



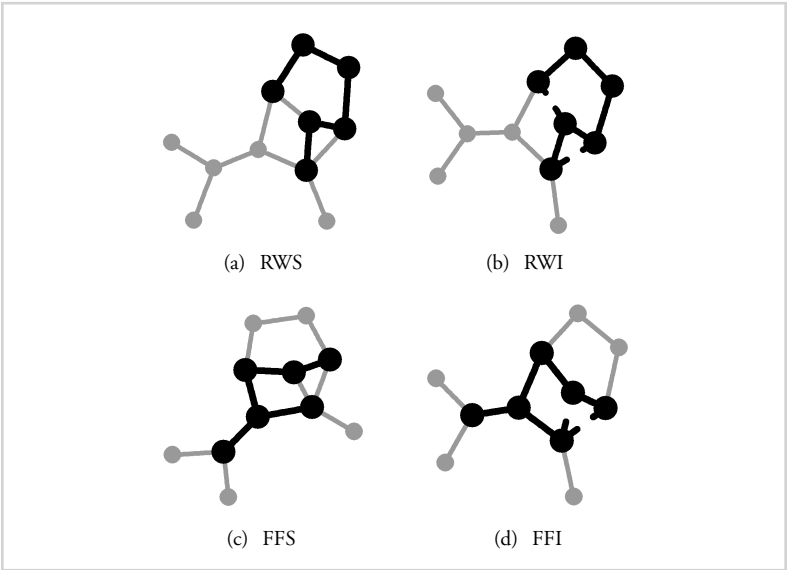
Slika 6.1

Delovanje pristopov zmanjševanja z vzorčenjem na primeru manjšega omrežja. Odebeljena črna vozlišča in povezave sestavljajo zmanjšana omrežja, dobljena z (a) naključnim izbiranjem vozlišč, (b) naključnim izbiranjem vozlišč glede na stopnjo, (c) naključnim izbiranjem povezav ter (d) naključnim izbiranjem povezav s korakom indukcije, kjer so odebeljene črne povezave izbrane naključno, črtkane pa dodane v zmanjšano omrežje pri koraku indukcije.

damo le določeno število povezav [40] (slika 6.2(c)). Število izbranih povezav je porazdeljeno geometrijsko s srednjo vrednostjo $p/(1-p)$, kjer p nastavimo na 0.7 [40]. Na vsakem koraku je tako v povprečju izbranih 2,33 povezav. FFS dobro ohrani porazdelitev vhodnih in izhodnih stopenj [40], slabo pa se izkaže pri ohranjanju nakopičenosti [184]. V oba pristopa, RNS in FFS, vpeljemo korak indukcije, dobljena pristopa imenujemo preiskovanje z naključnim sprehodom z indukcijo (angl. random walk sampling with subgraph induction, RWI) na sliki 6.2(b) in delno preiskovanje v širino z indukcijo (angl. forest-fire sampling with subgraph induction, FFI) na sliki 6.2(d). Po našem vedenju RWI v literaturi še ni bil analiziran, medtem ko FFI slabše kot RLI ohrani porazdelitev stopenj vozlišč in nakopičenosti [184].

6.2 Uporabljena omrežja

Delovanje pristopov primerjamo pri zmanjševanju 12 omrežij različnih tipov in velikosti (tabela 6.2). Tri omrežja spadajo v skupino omrežij sodelovanj, kjer vozlišča predstavljajo raziskovalce, ki so med seboj povezani, če so soav-



Slika 6.2

Delovanje pristopov zmanjševanja s preiskovanjem na primeru manjšega omrežja. Odebeljena črna vozlišča in povezave sestavljajo zmanjšana omrežja, dobljena s (a) preiskovanjem z naključnim sprehodom, (b) preiskovanja z naključnim sprehodom s korakom indukcije,(c) delnega preiskovanje v širino ter (d) delnega preiskovanje v širino s korakom indukcije. V obeh primerih zmanjševanja s korakom indukcije so odebeljene črne povezave izbrane naključno, črtkane pa dodane v zmanjšano omrežje pri koraku indukcije.

torji vsaj enega članka. Omrežje *ca-hep* je omrežje sodelovanj med raziskovalci na področju fizike osnovnih delcev, *ca-astro* na področju astrofizike ter omrežje *ca-dblp* na področju računalništva. Biološki omrežji *yeast* in *human* sta omrežji interakcij med beljakovinami *S. cerevisiae* oziroma *H. sapiens*. Vozlišča v obeh omrežjih predstavljajo beljakovine, povezave pa pomenijo interakcije med njimi. Omrežje *cit-hep* je omrežje citiranj med prispevki iz področja fizike visokih delcev. Vozlišča omrežja so prispevki, ki so povezani, če se med seboj citirajo. Omrežja *brightkite*, *slashdot* in *youtube* so omrežja prijateljstev istoimenskih spletnih portalov, kjer so vozlišča osebe, povezave pa pomenijo prijateljstvo med njimi. Omrežje *email* sestavljajo spletni naslovi Evropskega raziskovalnega inštituta, ki so med seboj povezani, če je bilo med njimi poslano vsaj eno elektronsko sporočilo. Omrežje *nd.edu* je omrežje povezav med spletnimi stranmi Univerze Notre Dame na domeni *nd.edu*, kjer so vozlišča spletne strani, povezave pa spletne povezave med njimi. Omrežje *flickr* je omrežje fotografij spletnega omrežja Flickr. Vozlišča predstavljajo fotografije, ki so povezane, če imajo skupne metapodatke, kot so na primer album, kjer je fotografija shranjena, lokacija posnetka ali avtor fotografije.

Tabela 6.2

V analizi uporabljena realna omrežja in njihove osnovne lastnosti.

Omrežje	Vozlišča	Povezave	Povprečna stopnja	Koeficient nakopičenosti	Gostota
yeast [197]	5.717	48.259	16,9	0,068	$2{,}9 \times 10^{-3}$
ca-hep [91]	12.008	237.010	39,5	0,660	$3{,}3 \times 10^{-3}$
human [197]	15.921	220.019	27,6	0,021	$1{,}7 \times 10^{-3}$
ca-astro [91]	18.772	396.160	42,2	0,318	$2{,}2 \times 10^{-3}$
cit-hep [86]	27.240	342.437	25,1	0,120	$9{,}2 \times 10^{-4}$
brightkite [101]	58.228	214.078	7,4	0,111	$1{,}3 \times 10^{-4}$
slashdot [103]	82.168	948.464	23,1	0,024	$2{,}8 \times 10^{-4}$
flickr [97]	105.938	2.316.948	43,7	0,402	$4{,}1 \times 10^{-4}$
email [91]	265.214	420.045	3,2	0,004	$1{,}2 \times 10^{-5}$
ca-dblp [92]	317.080	1.049.866	6,6	0,306	$2{,}1 \times 10^{-5}$
nd.edu [107]	325.729	1.497.134	9,1	0,097	$2{,}8 \times 10^{-5}$
youtube [198]	1.134.890	2.987.624	5,2	0,006	$4{,}5 \times 10^{-6}$

6.3 Analiza in rezultati

Omrežja, opisana v tabeli 6.2, z vsakim pristopom zmanjšamo na 30 različnih velikosti: 0,2–1 % velikosti osnovnega omrežja s korakom 0,2 % in 2–20 % velikosti osnovnega omrežja s korakom 2 %. Za vsako omrežje in za vsako velikost izvedemo 100 ponovitev zmanjševanja s posameznim pristopom ter rezultate povprečimo.

Osnovna in zmanjšana omrežja primerjamo na podlagi štirih lastnosti: porazdelitve stopenj vozlišč in nakopičenosti, povprečne stopnje vozlišč ter gostote omrežja. Porazdelitev stopenj vozlišč in nakopičenosti primerjamo z D-statistiko Kolmogorov-Smirnova, ki meri razdaljo med porazdelitvama posamezne lastnosti osnovnega in zmanjšanega omrežja. Pri povprečni stopnji vozlišč in gostoti primerjamo dejanske vrednosti posamezne lastnosti osnovnega in zmanjšanega omrežja.

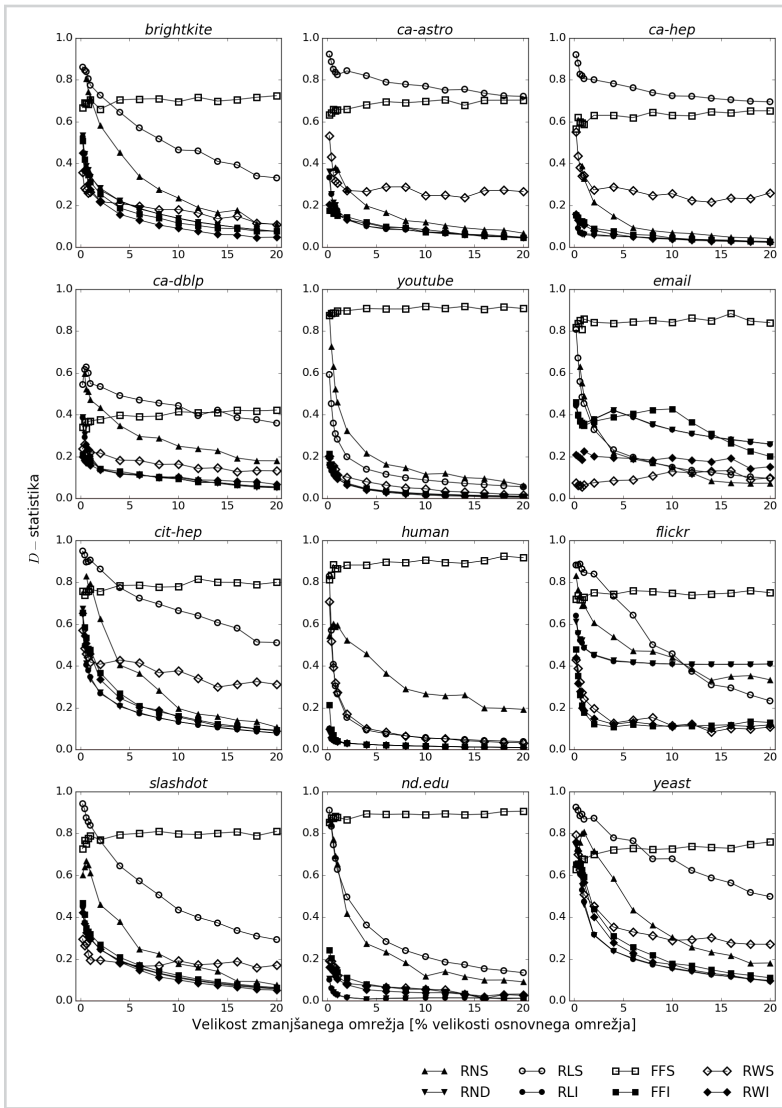
Rezultati primerjave porazdelitev stopenj vozlišč so prikazani na sliki 6.3. Izkaže se, da se porazdelitev stopenj bolje ohrani pri zmanjševanju z indukcijo kot brez nje. Z izjemo omrežja *email*, kjer je najboljši pristop RWS, pristopi z indukcijo izboljšajo delovanje pristopov brez indukcije. V splošnem porazdelitev stopenj vozlišč najbolje ohranijo RLI in RWI, najslabše pa FFS (tabela 6.3). Na zmanjšanih omrežjih velikosti manj kot 0,6 % osnovnih omrežij pristopi delujejo slabše, saj zmanjšana omrežja postajajo vse manj povezana, sestavljena iz večjega števila manjših komponent. Tako so manj podobna osnovnim

omrežjem, ne samo v porazdelitvi stopenj vozlišč, temveč tudi v ostalih analiziranih lastnostih.

Rezultati primerjave porazdelitve nakopičenosti so prikazani na sliki 6.4. V splošnem je porazdelitev nakopičenosti slabše ohranjena kot porazdelitev stopenj vozlišč, prav tako so razlike med uspešnostjo pristopov manjše. V splošnem se kot najboljša izkažeta RDS in RLI, najslabše pa nakopičenost ohrani RNS (tabela 6.3). Pri večini omrežij se uspešnost pristopov s korakom indukcije poveča, ne velja pa to za za omrežje *slashdot*, kjer nakopičenost najbolje ohrani FFS, ter za omrežji *youtube* in *email*, kjer slabo delujeta FFI oziroma RLI. Vsa tri omrežja imajo v primerjavi z ostalimi manjšo tranzitivnost (tabela 6.2). Možen razlog za slabše delovanje pristopov z indukcijo je tako izbira dodatnih povezav v zmanjšano omrežje, kar poveča tranzitivnost zmanjšanega omrežja in s tem tudi gostoto povezav v lokalni okolici vozlišč.

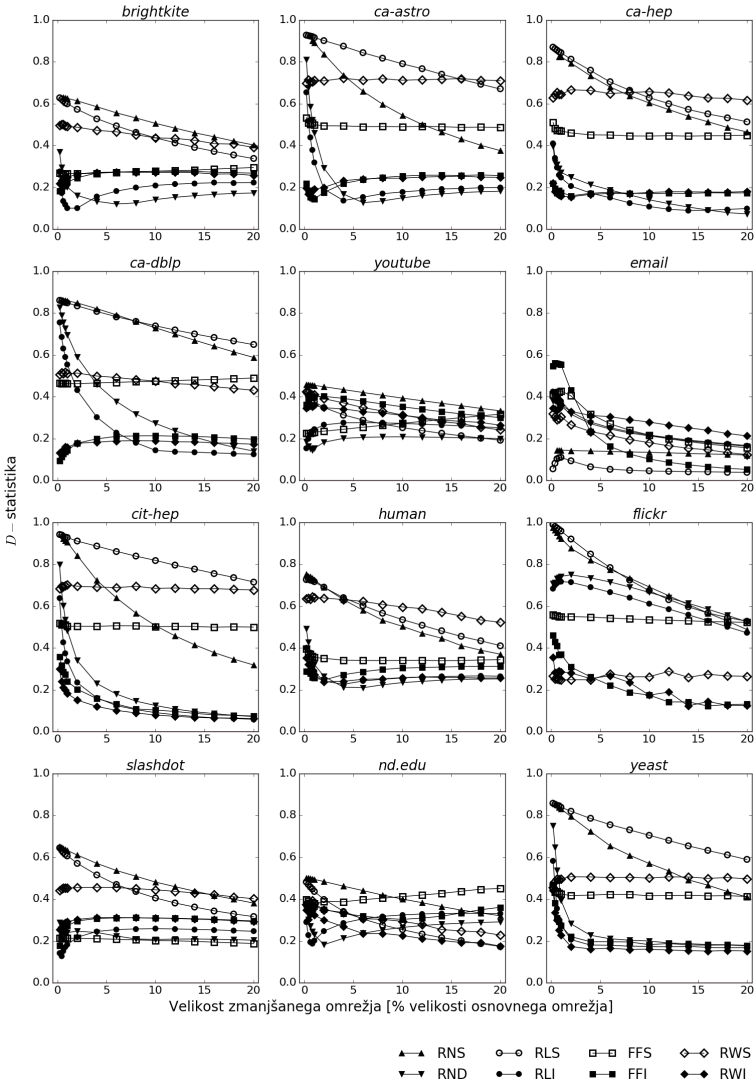
Povprečno stopnjo vozlišč v splošnem vsi pristopi ohranijo slabo (rezultati so prikazani na sliki 6.5). V povprečju najbolje delujeta RWS in FFS, najslabše pa RLI (tabela 6.3). Primerjava povprečne stopnje vozlišč kaže tudi najbolj izrazito razliko med pristopi z indukcijo in brez nje. Na povprečno stopnjo zmanjšanih omrežij bolj kot lastnosti omrežij vplivajo značilnosti pristopov za zmanjševanje. Zmanjšana omrežja z RNS in s pristopi brez indukcije imajo povprečno stopnjo pod 5. Pri zmanjševanju z RLS naključno vzorčimo povezave in njihova krajišča, kar ustvari redko in nepovezano zmanjšano omrežje z nizko povprečno stopnjo [184], podobno velja za RNS. Pri RWS se naključni sprehajalec redko vrne v že obiskano vozlišče, pri zmanjševanju s FFS in parametru p, nastavljenemu na $0,7$, po definiciji na vsakem koraku v povprečju izberemo $2,33$ povezav [40]. Temu primerno ima zmanjšano omrežje nižjo povprečno stopnjo. Po drugi strani pa pristopi s korakom indukcije ustvarijo omrežja z višjo povprečno stopnjo kot osnovna omrežja. Slednje ne velja le za omrežji *email* in *youtube*, ki imata najnižjo povprečno stopnjo med vsemi in jo pristopi z indukcijo bolje ohranijo pri večjih zmanjšanih omrežjih.

Rezultati primerjave gostote omrežij so prikazani na sliki 6.6, kjer je zaradi preglednosti v grafih uporabljena logaritemska skala. V poglavju 4 smo pokazali, da pri zmanjševanju z združevanjem obstaja potenčno razmerje med velikostjo omrežja in njegovo gostoto; velika omrežja so redka, z zmanjševanjem pa postajajo gostejša. Iz slike 6.6 opazimo, da podobno velja tudi za ostale pristope zmanjševanja z izjemo RNS. Le-ta ohranja gostoto osnovnih omrežij ne glede na velikost zmanjšanih omrežij, v nekaterih primerih omrežja postanejo gostejša pri velikostih pod 1 % osnovnih omrežij. Po drugi strani



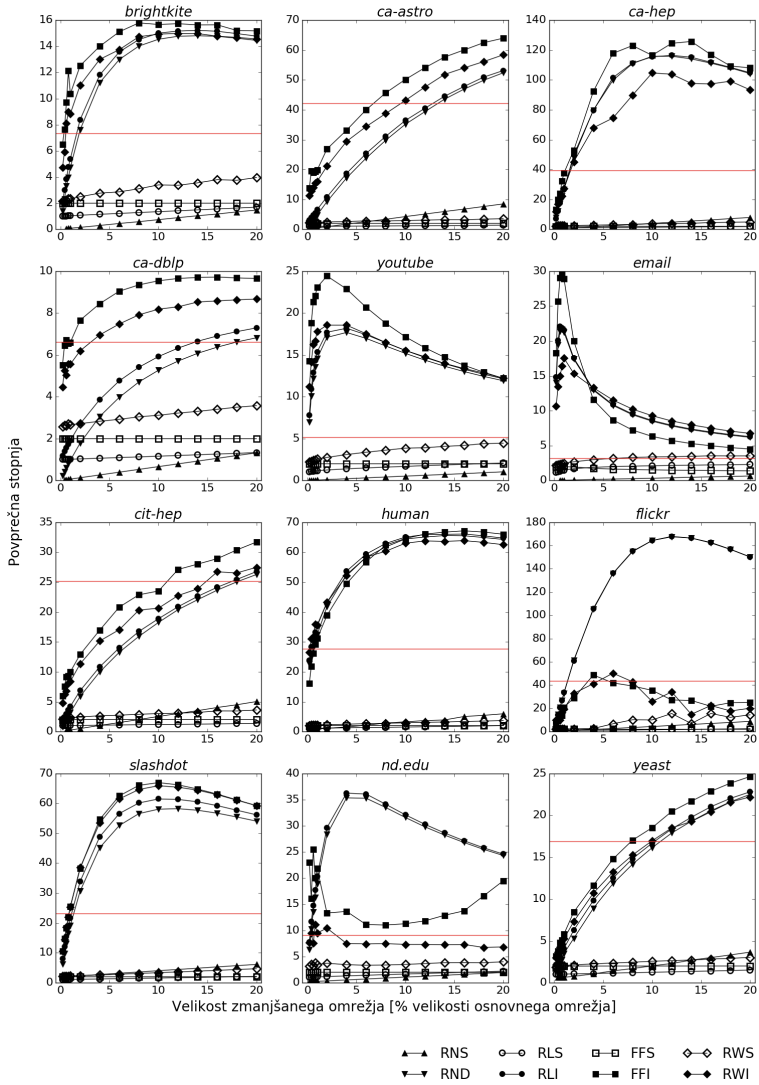
Slika 6.3

Primerjava porazdelitve stopenj vozlišč osnovnih in zmanjšanih omrežij.



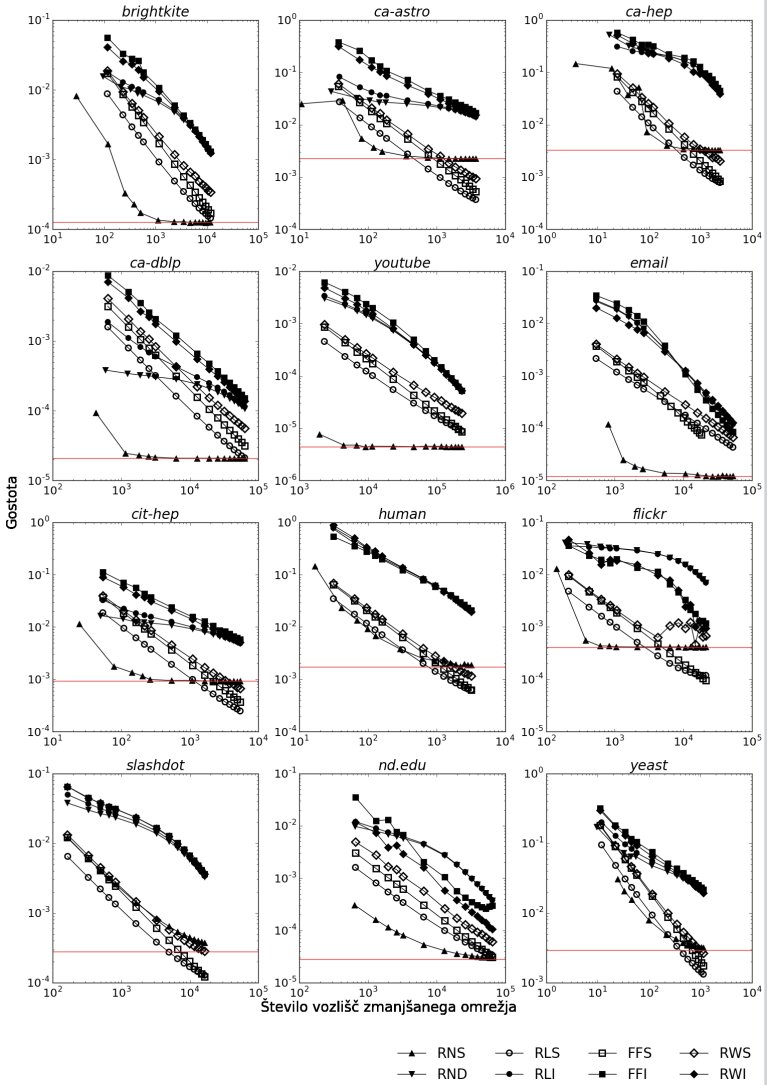
Slika 6.4

Primerjava porazdelitve nakopičenosti osnovnih in zmanjšanih omrežij.



Slika 6.5

Primerjava povprečne stopnje osnovnih in zmanjšanih omrežij. Povprečna stopnja osnovnih omrežij je označena z rdečo črto.



Slika 6.6

Primerjava gostote osnovnih in zmanjšanih omrežij. Skala na obeh oseh je logaritemska. Gostota osnovnih omrežij je označena z rdečo črto.

Tabela 6.3

Najboljša dva in najslabši pristop glede na mero A (glej poglavje 3) za ohranjanje posameznih lastnosti pri zmanjšanih omrežjih velikosti 10 % osnovnih omrežij.

Lastnost	Najboljši	Drugi najboljši	Najslabši
Porazdelitev stopenj vozlišč	RWI (0,18)	RLI (0,23)	FFS (0,96)
Porazdelitev nakopičenosti	RDS (0,26)	RLI (0,27)	RNS (0,81)
Povprečna stopnja	RWS (0,29)	RWI (0,43)	RLI (0,62)
Gostota	RNS (0,06)	FFS (0,24)	FFI (0,91)

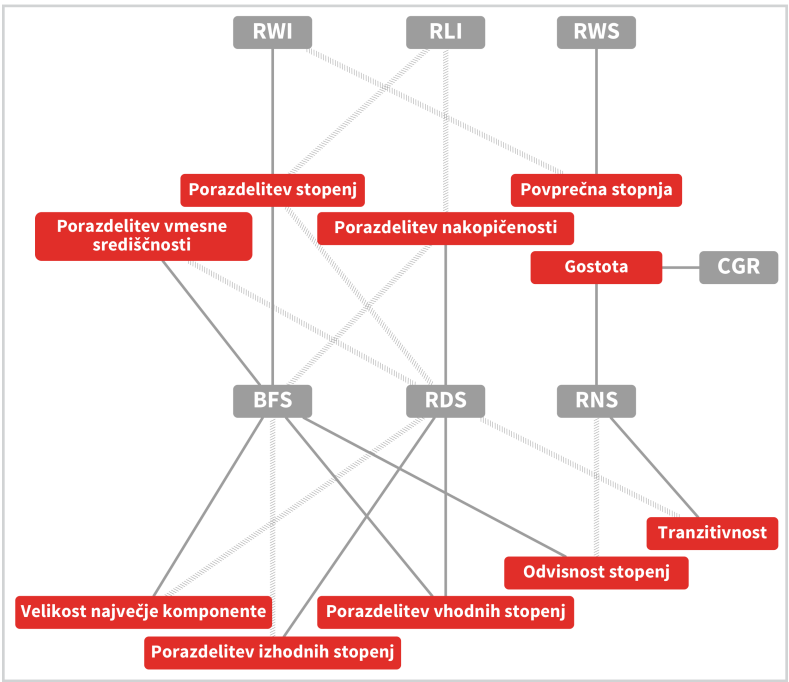
s pristopi zmanjševanja z indukcijo pričakovano dobimo gostejša omrežja kot pri pristopih brez indukcije, saj se gostota z dodajanjem povezav veča.

V poglavju 3 smo podrobneje analizirali uspešnost pristopov pri zmanjšanih omrežjih velikosti 10 % osnovnih omrežij. Za primerjavo in širši vpogled v delovanje pristopov v tabeli 6.3 predstavimo najboljše in najslabše pristope za ohranjanje posameznih lastnosti iz analize tega poglavja. Pristopi so razvrščeni glede na mero A (poglavje 3). Tabela prikazuje rezultate za zmanjšana omrežja velikosti 10 % osnovnih omrežij, razvrstitve pristopov so zelo podobne za velikosti med 0,6 % in 20 % osnovnih omrežij. Manjša omrežja, pod 0,6 % velikosti osnovnih omrežij, postajajo vse bolj nepovezana in vse manj podobna osnovnim omrežjem, kar vpliva na nižjo uspešnost zmanjševanja.

6.4 Shema za izbiro pristopa

Na podlagi rezultatov analiz sestavimo shemo za pomoč pri izbiri pristopa za zmanjševanje omrežja. Shema je predstavljena na sliki 6.7. Z rdečimi vozlišči so označene lastnosti omrežij, s sivimi pa pristopi za zmanjševanje. Za pristope so uporabljene kratice, opisane v tabeli 6.1. Polne povezave med vozlišči pomenijo najboljši pristop za ohranjanje posamezne lastnosti, črtkane pa drugo najboljšo izbiro (glej tabeli 3.5 in 6.3). Zgornja polovica sheme je sestavljena iz rezultatov primerjave pristopov analize tega poglavja, spodnja pa iz poglavja 3. Za ohranjanje gostote se kot najboljši izkaže pristop CGR (poglavje 4), zato v shemo poleg slednjega iz obeh analiz dodamo le najboljši pristop, ki je v obeh primerih RNS.

Shema je sestavljena iz rezultatov analiz zmanjšanih omrežjih velikosti 10 % osnovnih omrežij. Vanjo so vključene lastnosti, ki smo jih opazovali v analizah. Izmed pristopov za zmanjševanje pa se v shemi ne pojavijo vsi. Manjkajo RLS ter delna preiskovanja v širino, FFS in FFI. Pomanjkljivosti



Slika 6.7

Shema za izbiro pristopa za zmanjševanje glede na lastnosti, ki jih med zmanjševanjem želimo ohraniti. Siva vozlišča označujejo pristope za zmanjševanje, rdeča pa lastnosti, ki jih želimo z zmanjševanjem ohraniti. Polne povezave pomenijo najboljši, črtkane pa drugi najboljši pristop (glej poglavji 3 in 6).

RLS zelo dobro odpravimo z dodanim korakom indukcije pri pristopu RLI, ki je vključen v shemo. Podobno velja za delno preiskovanje v širino; s FFI sicer izboljšamo delovanje navadnega FFS, a ne dovolj, saj ostali pristopi vseeno lastnosti ohranijo bolje od preiskovanj v širino.

Poleg razlik med pristopi so različni tudi razlogi za uporabo zmanjševanja. Najpogosteje želimo veliko omrežje zmanjšati zaradi časovne ali prostorske zahtevnosti algoritmov, potrebnih za analizo. Nemalokrat je veliko omrežje nemogoče prikazati in ga zmanjšamo za namene učinkovitejšega prikaza. Pristop za zmanjševanje glede na omenjena dva cilja izberemo v treh korakih:

1. Določimo, kako veliko naj bo zmanjšano omrežje. Če velikost ni ključnega pomena, naj bo zmanjšano omrežje velikosti med 1 % in 10 % osnovnega omrežja.

2. Določimo, katere lastnosti omrežja so pomembne pri analizi ali prikazu in

želimo, da se med zmanjševanjem ohranijo.

3. Pristop za zmanjševanje izberemo glede na shemo na sliki 6.7.

Primeri uporabe sheme:

- če želimo prikazati, kako gosto je omrežje, izberemo CGR. Ker pa je slednji pristop časovno zahteven na zelo velikih omrežjih, v tem primeru raje izberemo RNS;
- porazdelitev vhodnih stopenj enako dobro ohranita tako BFS kot RDS, zato pri izbiri pristopa upoštevamo še uspešnost ohranjanja drugih lastnosti:
- če želimo prikazati povezanost omrežja v smislu nakopičenosti, izberemo RDS. Če hkrati želimo, da se dobro ohrani tudi porazdelitev stopenj in odvisnost med stopnjami, je boljša izbira BFS.

6.5 Sklepne ugotovitve

Ahmed in sodelavci [59] so pokazali, da se z izbiro dodatnih povezav v zmanjšano omrežje pri zmanjševanju z naključnim izbiranjem povezav izboljša delovanje osnovnega naključnega izbiranja. V tem poglavju smo raziskali, če lahko na enak način izboljšamo pristope zmanjševanja s preiskovanjem. Opazovali smo delovanje delnega preiskovanja v širino in preiskovanja z naključnim sprehodom s korakom indukcije. Primerjali smo ju z več pristopi na 12 realnih omrežjih različnih tipov in velikosti. Opazovali smo, kako se pri različnih velikostih zmanjšanih omrežij spreminjajo lastnosti, kot so porazdelitev stopenj vozlišč in nakopičenosti, povprečna stopnja ter gostota.

Pokazali smo, da korak indukcije izboljša delovanje tako naključnega izbiranja povezav kot tudi delnega preiskovanja v širino in preiskovanja z naključnim sprehodom pri ohranjanju porazdelitve stopenj vozlišč in nakopičenosti. Na nekaj omrežjih so rezultati nakazali, da na ohranjanje podobnosti med osnovnim in zmanjšanim omrežjem vplivajo lastnosti osnovnega omrežja. Nakopičenost omrežij z nižjo tranzitivnostjo bolje ohranijo pristopi brez indukcije. Zanesljivejši dokaz slednjega sodi med možnosti nadaljnjega dela, saj bi za potrditev statistično značilnih razlik med uspešnostjo pristopov v odvisnosti od lastnosti osnovnih omrežij potrebovali večjo množico omrežij. Nasprotno pa lastnosti osnovnega omrežja manj vplivajo na ohranjanje

povprečne stopnje, saj se izkaže, da je povprečna stopnja zmanjšanih omrežij bolj odvisna od značilnosti pristopov za zmanjševanje. Zmanjšana omrežja z naključnim izbiranjem vozlišč ter s pristopi brez indukcije imajo povprečno stopnjo pod 5. Po drugi strani pa imajo omrežja, zmanjšana s pristopi z indukcijo, višjo povprečno stopnjo kot osnovna omrežja. V splošnem se povprečna stopnja omrežij ohrani slabo; rešitev za to bi lahko v nadaljnjih analizah iskali v pristopih z delno indukcijo, kjer bi pri koraku indukcije v zmanjšano omrežje dodali le delež povezav osnovnega omrežja. Delovanje pristopov z indukcijo in brez nje se razlikuje tudi pri ohranjanju gostote. Zmanjševanje z indukcijo ustvari gostejša omrežja, kar je pričakovano, saj vsebujejo dodatne povezave v primerjavi s pristopi brez indukcije. Razmerje med velikostjo in gostoto omrežij ohranijo vsi pristopi razen naključnega izbiranja vozlišč, ki ohrani gostoto omrežja ne glede na velikost pri zmanjšanih omrežjih nad 0,6 % velikosti osnovnih omrežij. Ta velikost pa se v splošnem izkaže za mejo, kjer se uspešnost pristopov poslabša. V zadnjem delu analize opazujemo primernost pristopov za ohranjanje posameznih lastnosti pri velikosti 10 % zmanjšanih omrežij. Povprečno stopnjo najbolje ohrani preiskovanje z naključnim sprehodom, z uporabo indukcije pa najbolje ohranimo porazdelitev stopenj. Gostoto najbolje ohrani naključno izbiranje vozlišč, če pa vozlišča izberemo glede na njihovo stopnjo, se najbolje ohrani nakopičenost. Za konec na podlagi vseh rezultatov sestavimo shemo, ki pomaga pri izbiri pristopa za zmanjševanje izbranega omrežja. V shemi izstopata preiskovanje v širino ter naključno izbiranje vozlišč glede na stopnjo, ki najbolje ohranita večino lastnosti osnovnih omrežij.

Zaključek

7

Zmanjševanje omrežij je proces, ki nam olajša razumevanje delovanja in lastnosti omrežij realnega sveta. Zmanjšana omrežja omogočajo hitrejšo analizo in učinkovitejši prikaz velikih omrežij. Podatki o analiziranem omrežju so lahko manjkajoči ali skriti. V tem primeru nas zanimajo razlike med nepopolnim in celotnim omrežjem, ki jih opazujemo pri spreminjanju omrežij med zmanjševanjem. Raziskovalci so predlagali veliko pristopov za zmanjševanje, sprva predvsem z namenom učinkovitejšega shranjevanja velikih omrežij [46, 47]. Zadnja leta pa se zmanjševanje uporablja za pomoč pri analizi in razumevanju vse večjih omrežij [48, 49, 51]. Nekateri pristopi za zmanjševanje so tako primerni za ohranjanje posameznih lastnosti [54–56], drugi so namenjeni določenim vrstam omrežij [45, 52, 53]. Bistvenega pomena pri vseh pa je, da z zmanjševanjem ohranijo podobnost med osnovnim in zmanjšanim omrežjem.

V disertaciji se ukvarjamo z analizo spreminjanja omrežij med zmanjševanjem in primerjavo pristopov za zmanjševanje. Osrednja tri poglavja disertacije sestavljajo trije objavljeni članki [38, 60, 61]. V prvem članku obravnavamo vpliv tipa in velikosti osnovnih omrežij ter velikost zmanjšanih omrežij na ohranjanje lastnosti med zmanjševanjem. Predlagamo mero za oceno uspešnosti zmanjševanja, s katero primerjamo pristope med seboj. Izkaže se, da večja zmanjšana omrežja lastnosti ohranijo bolje, pri velikostih 10 % osnovnih omrežij pa dosežemo kompromis med velikostjo in ohranjanjem podobnosti z osnovnim omrežjem. Nasprotno pa tip in velikost osnovnih omrežij na uspešnost zmanjševanja ne vplivata. S predlagano mero primerjamo pristope glede na uspešnost ohranjanja posameznih lastnosti pri zmanjšanih omrežjih velikosti 10 % osnovnih omrežij. V splošnem se najbolje izkažeta preiskovanje v širino in naključno izbiranje vozlišč glede na stopnjo.

V drugih dveh člankih obravnavamo spreminjanje omrežij med zmanjševanjem. V drugem članku opazujemo, kako se pri zmanjševanju z združevanjem spreminja gostota omrežij. Dokažemo potenčno razmerje med velikostjo omrežij in njihovo gostoto. Razmerje velja za osnovna omrežja [24], močnejše pa postane z upoštevanjem tudi zmanjšanih omrežij. V tretjem članku pokažemo, da se povezovanje skupin vozlišč pri zmanjševanju družbenih in informacijskih omrežij spreminja. Osnovna družbena omrežja vsebujejo gosto povezane skupine vozlišč, redko povezane med seboj. Skupine v informacijskih omrežjih so redkejše in gosteje povezane med seboj. Pri zmanjševanju z različnimi pristopi v obeh primerih skupine postanejo gosteje povezane znotraj skupine in redkeje povezane med seboj.

V dodatnem poglavju disertacije medsebojno primerjamo pristope zmanj-

ševanja z vzorčenjem in s preiskovanjem. V študijah se predvsem zmanjševanje z naključnim izbiranjem povezav izkaže kot slabši pristop, saj ustvari redka in nepovezana zmanjšana omrežja, ki slabo ohranijo lastnosti osnovnih [40]. Z vpeljavo indukcije, kjer v zmanjšano omrežje dodamo povezave iz osnovnega omrežja, se izboljša delovanje naključnega izbiranja povezav [59]. Na podlagi slednje študije vpeljemo indukcijo v pristopa zmanjševanja s preiskovanjem in analiziramo, kako se pri tem spremeni uspešnost zmanjševanja. Analiza pokaže, da pristopi z indukcijo izboljšajo delovanje pristopov brez indukcije pri ohranjanju porazdelitve stopenj vozlišč in nakopičenosti. Ker z indukcijo v omrežje dodajamo povezave, so zmanjšana omrežja gostejša z višjo povprečno stopnjo vozlišč. Rezultati pokažejo, da pristopi zmanjševanja z vzorčenjem ne delujejo vedno slabše od zmanjševanj s preiskovanjem, opaznejša razlika pri ohranjanju lastnosti pa je med pristopi z indukcijo in brez nje.

Pri pripravi dispozicije doktorske disertacije smo postavili hipotezo, da so si omrežja enega tipa dovolj podobna, da jih lahko učinkovito zmanjšamo z enakimi pristopi. Za na primer družbena omrežja je lahko primeren nek pristop, ki na informacijskih omrežjih ne deluje dobro. Na podlagi hipoteze smo kot enega izmed prispevkov disertacije navedli algoritem za prilagojeno zmanjševanje omrežij. Algoritem bi prilagodil zmanjševanje lastnostim velikega omrežja in ga zmanjšal tako, da bi se lastnosti čimbolj ohranile. Na množici realnih omrežij in pristopih, ki smo jih uporabili pri analizah, se predpostavke izkažejo za napačne. Uspešnost zmanjševanja ni toliko odvisna od lastnosti omrežja, ki ga zmanjšujemo, temveč bolj od pristopa, ki ga uporabimo. Opomnimo, da smo v analizo zajeli raznoliko množico pristopov in omrežja različnih tipov, velikosti in lastnosti, ki se uporabljajo v podobnih študijah. Kljub temu bi lahko bili rezultati drugačni z izbiro drugih omrežij, lastnosti in pristopov.

Če povzamemo, izkaže se, da lastnosti osnovnega omrežja nimajo velikega vpliva na uspeh zmanjševanja. Lahko pa iz rezultatov sklepamo na razlike med pristopi za zmanjševanje glede na velikost zmanjšanih omrežij in lastnosti, ki jih želimo ohraniti. Vsi pristopi, razen zmanjševanj z združevanjem, omogočajo nastavljanje velikosti osnovnega omrežja, na primer na določeno število vozlišč ali določen odstotek velikosti osnovnega omrežja. Analiza pokaže, da se lastnosti osnovnega omrežja bolje ohranijo pri večjih zmanjšanih omrežjih. Pogosteje pa uporabimo pristope za zmanjševanje, ker želimo omrežje čimbolj zmanjšati. Pri velikostih zmanjšanih omrežij med 1 in 10 % osnovnega omrežja dosežemo kompromis med velikostjo in podobnostjo z osnov-

nim omrežjem. Glede na podobne študije [40, 57, 63] in naše analize, smo za podrobnejšo primerjavo pristopov predlagali zmanjšana omrežja velikosti 10 % osnovnih omrežij. Za to velikost oblikujemo shemo, ki pomaga pri izbiri pristopa za zmanjševanje izbranega omrežja. V shemi izberemo pristop za zmanjševanje omrežja glede na lastnosti, ki želimo, da se med zmanjševanjem ohranijo. V splošnem kot najboljša pristopa za ohranjanje posameznih lastnosti izstopata preiskovanje v širino in naključno izbiranje vozlišč glede na stopnjo. Namesto algoritma, ki bi omrežju prilagodil zmanjševanje, tako izbiro pristopa prilagodimo ciljem zmanjševanja.

Izvirni prispevki disertacije so trije:

- mera za ceno uspešnosti zmanjševanja, s katero primerjamo različne pristope in analiziramo njihovo uspešnost pri ohranjanju različnih lastnosti omrežij med zmanjševanjem,
- analiza spreminjanja omrežij med zmanjševanjem, kjer podrobneje opazujemo spreminjanje gostote in povezovanje skupin vozlišč med zmanjševanjem,
- shema za izbiro pristopa za zmanjševanje, ki pomaga pri izbiri primernega pristopa glede na lastnosti, ki jih želimo med zmanjševanjem ohraniti.

Našteti prispevki pomenijo pomemben doprinos na področju zmanjševanja omrežij. Rezultati namreč prispevajo k razumevanju spreminjanja omrežij med zmanjševanjem z možno uporabo pri hitrejši analizi in učinkovitejšemu prikazu velikih omrežij.

Študija je nakazala več možnosti nadaljnjih raziskav. Analiza večje množice realnih omrežij bi dokazala ali ovrgla vpliv lastnosti osnovnih omrežij na uspešnost zmanjševanja. Primer slednjega zaznamo pri nakopičenosti, ki se bolje ohrani s pristopi brez indukcije na omrežjih z nižjo tranzitivnostjo. Prav tako bi bila uporabna analiza časovne in prostorske zahtevnosti pristopov za zmanjševanje. Opazili smo, da so med njimi velike razlike zlasti v časovni zahtevnosti zmanjševanja. Podrobnejša primerjava časovne in prostorske zahtevnosti pristopov bi ponudila kriterij več za pomoč pri izbiri pristopa za zmanjševanje. Izboljšala bi se lahko tudi predlagana mera za oceno uspešnosti zmanjševanja. Primernejša bi bila mera, ki bi pristope ocenila neodvisno od analizirane množice omrežij in lastnosti. Kar nekaj pa je tudi možnosti

za izboljšave posameznih pristopov. Na primer zmanjšana omrežja s pristopi z indukcijo imajo povprečno stopnjo precej višjo kot osnovna omrežja, pristopi brez indukcije pa delujejo ravno obratno. Kombiniranje pristopov bi lahko izboljšalo uspešnost pristopov pri ohranjanju lastnosti. Ena možnost je delna indukcija, kjer bi v zmanjšano omrežje dodali le delež povezav osnovnega omrežja namesto vseh, druga možnost pa je hibridno preiskovanje, kjer bi omrežje preiskovali hkrati iz različnih naključno izbranih začetnih vozlišč.

LITERATURA

[1] M. E. J. Newman. *Networks: An introduction*. Oxford University Press, 2009.

[2] R. Cohen and S. Havlin. *Complex networks: structure, robustness and function*. Cambridge University Press, 2010.

[3] J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 290. Macmillan London, 1976.

[4] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[5] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[6] L. Šubelj, Š. Furlan, and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, 38(1):1039–1052, 2011.

[7] K. T. S. Oldham. *The Doctrine of Description: Gustav Kirchhoff, Classical Physics, and the" purpose of All Science" in 19th-century Germany*. ProQuest, 2008.

[8] R. Shields. Cultural topology: The seven bridges of königsburg, 1736. *Theory, Culture & Society*, 29(4-5):43–57, 2012.

[9] T. R. Jensen and B. Toft. *Graph coloring problems*, volume 39. John Wiley & Sons, 2011.

[10] W. W. Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.*, 33(4):452–473, 1977.

[11] B. Bollobás. *Random graphs*. Springer, 1998.

[12] S. Janson, T. Luczak, and A. Rucinski. *Random graphs*, volume 45. John Wiley & Sons, 2011.

[13] S. Strogatz. Exploring complex networks. nature, 410, 268–276. *Nonlinear Dynamics*, 2001.

[14] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[15] M. E. J. Newman. 2 random graphs as models of networks. *Handbook of Graphs and Networks: From the Genome to the Internet*, 2006.

[16] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[17] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2): 167–256, 2003. ISSN 0036-1445.

[18] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[19] Slovenske železnice. Zemljevid prog. http://www.slo-zeleznice.si/sl/potniki/vozni-redi/zemljevid-prog, 2015.

[20] B. Cheswick, H. Burch, and S. Branigan. Mapping and visualizing the internet. In *USENIX Annual Technical Conference, General Track*, pages 1–12. Citeseer, 2000.

[21] N. Blagus and M. Bajec. Omrežje sodelovanj med avtorji prispevkov iz Informatice in Uporabne informatike. *Uporabna informatika*, 23(1):22–31, 2015.

[22] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[23] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM, 2012.

[24] P. J. Laurienti, K. E. Joyce, Q. K. Telesford, J. H. Burdette, and S. Hayasaka. Universal fractal scaling of self-organized networks. *Physica A*, 390(20):3608–3613, 2011.

[25] J. Scott and P. J. Carrington. *The SAGE handbook of social network analysis*. SAGE publications, 2011.

[26] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *P. Natl. Acad. Sci. USA*, 105(4):1118–1123, 2008.

[27] M. Girvan and M. E. J Newman. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.

[28] C. Song, S. Havlin, and H. A. Makse. Self-similarity of complex networks. *Nature*, 433 (7024):392–395, 2005.

[29] M. A. Porter, J. P. Onnela, and P. J. Mucha. Communities in networks. *Not. Am. Math. Soc.*, 56(9):1082–1097, 2009.

[30] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174, 2010.

[31] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.*, 96(11):114102, 2006.

[32] M. E. J. Newman and E. A Leicht. Mixture models and exploratory analysis in networks. *P. Natl. Acad. Sci. USA*, 104(23):9564, 2007.

[33] S. Pinkert, J. Schultz, and J. Reichardt. Protein interaction networks—more than mere modules. *PLoS Computational Biology*, 6(1): e1000659, 2010.

[34] L. Šubelj and M. Bajec. Ubiquitousness of link-density and link-pattern communities in real-world networks. *Eur. Phys. J. B*, 85(1): 1–11, 2012.

[35] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1): 35–41, 1977.

[36] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67(2):026126, 2003.

[37] L. K. Gallos, C. Song, and H. A. Makse. A review of fractality and self-similarity in complex networks. *Physica A*, 386(2):686–691, 2007.

[38] N. Blagus, L. Šubelj, and M. Bajec. Self-similar scaling of density in complex real-world networks. *Physica A*, 391(8):2794–2802, 2012.

[39] W. X. Zhou, Z. Q. Jiang, and D. Sornette. Exploring self-similarity of complex cellular networks: The edge-covering method with simulated annealing and log-periodic sampling. *Physica A*, 375(2):741–752, 2007.

[40] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International conference on Knowledge Discovery and Data Mining*, pages 631–636. ACM, 2006.

[41] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS. In *Proceedings of the 22nd International Teletraffic Congress*, pages 1–8. IEEE, 2010.

[42] M. Zhong and K. Shen. Random walk based node sampling in self-organizing networks. *ACM SIGOPS Operating Systems Review*, 40 (3):49–55, 2006.

[43] M. Hamann, G. Lindner, H. Meyerhenke, C. L. Staudt, and D. Wagner. Structure-preserving sparsification methods for social networks. *e-print arXiv:1601.00286*, 2016.

[44] A. Rezvanian and M. R. Meybodi. Sampling social networks using shortest paths. *Physica A*, 424:254–268, 2015.

[45] Z. S. Jalali, A. Rezvanian, and M. R. Meybodi. Social network sampling using spanning trees. *Int. J. Mod. Phys. C*, 2015.

[46] N. Deo and B. Litow. A structural approach to graph compression. In *Proceedings of the 23th MFCS Workshop on Communications*, pages 91–101. Citeseer, 1998.

[47] M. Adler and M. Mitzenmacher. Towards compressing web graphs. In *Proceedings of the Data Compression Conference*, pages 203–212. IEEE, 2001.

[48] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J-H. Cui, and A. G. Percus. Reducing large internet topologies for faster simulations. In *Proceedings of the 4th International IFIP-TC6 Networking Conference*, pages 328–341. Springer, 2005.

[49] Chris Bennett. More efficient classification of web content using graph sampling. In *IEEE Symposium on Computational Intelligence and Data Mining*, pages 485–490. IEEE, 2007.

[50] D. Rafiei. Effectively visualizing large networks through sampling. In *Visualization*, pages 375–382. IEEE, 2005.

[51] D. Hennessey, D. Brooks, A. Fridman, and D. Breen. A simplification algorithm for visualizing the structure of complex graphs. In *Proceedings of the 12th International Conference on Information Visualisation*, pages 616–625. IEEE, 2008.

[52] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *P. Natl. Acad. Sci. USA*, 102(12):4221–4224, 2005.

[53] S.-W. Son, C. Christensen, G. Bizhani, D. V. Foster, P. Grassberger, and M. Paczuski. Sampling properties of directed networks. *Phys. Rev. E*, 86(4):046104, 2012.

[54] M. P. H. Stumpf and C. Wiuf. Sampling properties of random graphs: the degree distribution. *Phys. Rev. E*, 72(3):036118, 2005.

[55] F. Zhou, S. Malher, and H. Toivonen. Network simplification with minimal loss of connectivity. In *Proceedings of the 10th International Conference on Data Mining*, pages 659–668. IEEE, 2010.

[56] H. Maserrat and J. Pei. Community preserving lossy compression of social networks. In *Proceedings of the 12th International Conference on Data Mining*, pages 509–518. IEEE, 2012.

[57] S. H. Lee, P. J. Kim, and H. Jeong. Statistical properties of sampled networks. *Phys. Rev. E*, 73(1):016102, 2006.

[58] H. Sethu and X. Chu. A new algorithm for extracting a small representative subgraph from a very large graph. *e-print arXiv:1207.4825*, 2012.

[59] N. K. Ahmed, J. Neville, and R. R. Kompella. Network sampling via edge-based node selection with graph induction. Technical report, Purdue University, 2011.

[60] N. Blagus, L. Šubelj, and M. Bajec. Assessing the effectiveness of real-world network simplification. *Physica A*, 413:134–146, 2014.

[61] N. Blagus, L. Šubelj, G. Weiss, and M. Bajec. Sampling promotes community structure in social and information networks. *Physica A*, 432:206–215, 2015.

[62] C. Hübler, H. P. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis algorithms for representative subgraph sampling. In *Proceedings of the 8th International Conference on Data Mining*, pages 283–292. IEEE, 2008.

[63] C. Doerr and N. Blenn. Metric convergence in social network sampling. In *Proceedings of the 5th ACM workshop on HotPlanet*, pages 45–50. ACM, 2013.

[64] H. Toivonen, F. Zhou, A. Hartikainen, and A. Hinkka. Compression of weighted graphs. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 965–973. ACM, 2011.

[65] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, 2002.

[66] S. Wasserman. *Social network analysis: Methods and applications*. Cambridge university press, 1994.

[67] U. Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol.*, 25(2): 163–177, 2001.

[68] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68(6):065103, 2003.

[69] J. S. Kim, K. I. Goh, B. Kahng, and D. Kim. Fractality and self-similarity in scale-free networks. *New J. Phys.*, 9(6):177, 2007.

[70] F. Radicchi, J. J. Ramasco, A. Barrat, and S. Fortunato. Complex networks renormalization: Flows and fixed points. *Phys. Rev. Lett.*, 101(14):148701, 2008.

[71] M. Salehi, H. R. Rabiee, and A. Rajabi. Sampling from complex networks with high community structures. *Chaos*, 22(2):023126, 2012.

[72] A. S Maiya and T. Y. Berger-Wolf. Sampling community structure. In *Proceedings of the 19th international conference on World wide web*, pages 701–710. ACM, 2010.

[73] L. Šubelj, N. Blagus, and M. Bajec. Group extraction for real-world networks: The case of communities, modules, and hubs and spokes. In *Proceedings of the International Conference on Network Science*, pages 152–153, 2013.

[74] G. Weiss and L. Šubelj. nets-nodegroups v1.0. http://dx.doi.org/10.5281/zenodo.11589, 2014.

[75] M. E. J. Newman, A.-L. Barabási, and D. J. Watts. *The structure and dynamics of networks*. Princeton University Press, 2006.

[76] T. Feder and R. Motwani. Clique partitions, graph compression and speeding-up algorithms. In *Proceedings of the 23th Annual ACM Symposium on Theory of Computing*, pages 123–133. ACM, 1991.

[77] P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized blockmodeling*. Cambridge University Press, 2005.

[78] M. Kudelka, Z. Horak, V. Snasel, and A. Abraham. Social network reduction based on stability. In *Proceedings of the International Conference on Computational Aspects of Social Networks*, pages 509–514. IEEE, 2010.

[79] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, pages 27–40, 2006.

[80] T. Biedl, B. Brejová, and T. Vinař. Simplifying flow networks. *Lect. Notes Comput. Sc.*, pages 192–201, 2000.

[81] D. Gfeller and P. De Los Rios. Spectral coarse graining of complex networks. *Phys. Rev. Lett.*, 99(3):38701, 2007.

[82] J. Illenberger and G. Flötteröd. Estimating properties from snowball sampled networks. Technical report, VSP Working Paper 11-01, TU Berlin, Transport Systems Planning and Transport Telematics, 2011.

[83] H. Toivonen, F. Zhou, A. Hartikainen, and A. Hinkka. Network compression by node and edge mergers. In *Bisociative Knowledge Discovery*, pages 199–217. Springer, 2012.

[84] Y. Zhou, H. Cheng, and J.X. Yu. Graph clustering based on structural/attribute similarities. *Proc. VLDB Endowment*, 2(1): 718–729, 2009.

[85] L. Šubelj and M. Bajec. Robust network community detection using balanced propagation. *Eur. Phys. J. B*, 81(3):353–362, 2011.

[86] KDD Cup '03. http://www.cs.cornell.edu/projects/kddcup/, 2013.

[87] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 177–187. ACM, 2005.

[88] B. H. Hall, A. B. Jaffe, and M. Tratjenberg. The NBER patent citation data file, 2001.

[89] K. D Bollacker, S. Lawrence, and C. L. Giles. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the 2nd International Conference on Autonomous Agents*, pages 116–123. ACM, 1998.

[90] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Phys. Rev. E*, 70(5):056122, 2004.

[91] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):1–40, 2007.

[92] J. Yang and J. Leskovec. Community-affiliation graph model for overlapping network community detection. In *Proceedings of the 12th International Conference on Data Mining*, pages 1170–1175. IEEE, 2012.

[93] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Social synchrony: Predicting mimicry of user actions in online social media. In *International Conference on Computational Science and Engineering*, pages 151–158. IEEE, 2009.

[94] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 3. ACM, 2012.

[95] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pages 37–42. ACM, 2009.

[96] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):1–39, 2007.

[97] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, pages 548–556, 2012.

[98] M. E. J. Newman. Network data. http://www-personal.umich.edu/~mejn/netdata/, 2013.

[99] V. Batagelj, A. Mrvar, and M. Zaveršnik. *Network analysis of texts*. Univ. of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Dep. of Theoretical Computer Science, 2002.

[100] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings of the International Conference on Human Factors in Computing Systems*, pages 1361–1370. ACM, 2010.

[101] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1082–1090. ACM, 2011.

[102] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the 2nd International Semantic Web Conference*, pages 351–368. Springer, 2003.

[103] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Math.*, 6(1):29–123, 2009.

[104] S. Maniu, T. Abdessalem, and B. Cautis. Casting a web of trust over wikipedia: an interaction-based approach. In *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 87–88. ACM, 2011.

[105] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5): 604–632, 1999.

[106] G. Palla, I. J. Farkas, P. Pollner, I. Derenyi, and T. Vicsek. Directed network modules. *New J. Phys.*, 9(6):186, 2007.

[107] R. Albert, H. Jeong, and A.-L. Barabási. The diameter of the World Wide Web. *Nature*, 401:130–131, 1999.

[108] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, and Y. Yu. Zhishi. me-weaving chinese linking open data. In *Proceedings of the 10th International Semantic Web Conference*, pages 205–220. Springer, 2011.

[109] A. Clauset and C. Moore. Accuracy and scaling phenomena in internet mapping. *Phys. Rev. Lett.*, 94(1):018701, 2005.

[110] F. Viger, A. Barrat, L. Dall'Asta, C.-H. Zhang, and E. D. Kolaczyk. What is the real size of a sampled network? the case of the internet. *Phys. Rev. E*, 75(5):056111, 2007.

[111] M. Á. Serrano, M. Boguñá, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *P. Natl. Acad. Sci. USA*, 106(16):6483–6488, 2009.

[112] J.-P. Onnela, D. J. Fenn, S. Reid, M. A. Porter, P. J. Mucha, M. D. Fricker, and N. S. Jones. Taxonomies of networks from community structure. *Phys. Rev. E*, 86(3): 036104, 2012.

[113] L. K. Gallos and N. H. Fefferman. Revealing effective classifiers through network comparison. *e-print arXiv:1403.2668*, 2014.

[114] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *Comput. Commun. Rev.*, 29(4): 251–262, 1999.

[115] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

[116] S. Nadiv Soffer and A. Vázquez. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E*, 71(5): 057101, 2005.

[117] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[118] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85(21): 4626–4628, 2000.

[119] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67(2):026112, 2003.

[120] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2001.

[121] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307): 761–764, 2010.

[122] Mark E. J. Newman. The physics of networks. *Phys. Today*, 61(11):33–38, 2008.

[123] S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, and U. Alon. Coarse-graining and self-dissimilarity of complex networks. *Phys. Rev. E*, 71(1):016127, 2005.

[124] T. Carletti and S. Righi. Weighted fractal networks. *Physica A*, 389(10):2134–2142, 2010.

[125] B. B. Mandelbrot. *The fractal geometry of nature*. W. H. Freeman, 1983.

[126] A. Bunde and S. Havlin. *Fractals and disordered systems*. Springer-Verlag, Heidelberg, 1996.

[127] D. Ben-Avraham and S. Havlin. *Diffusion and reactions in fractals and disordered systems*. Cambridge University Press, Cambridge, 2000.

[128] K. I. Goh, G. Salvi, B. Kahng, and D. Kim. Skeleton and fractal scaling in complex networks. *Phys. Rev. Lett.*, 96(1):18701, 2006.

[129] C. Song, S. Havlin, and H. A. Makse. Origins of fractality in the growth of complex networks. *Nat. Phys.*, 2(4):275–281, 2006.

[130] H. D. Rozenfeld, L. K. Gallos, C. Song, and H. A. Makse. Fractal and transfractal scale-free networks. *e-print arXiv:0808.2206v1*, 2008.

[131] A. L. Barabási, E. Ravasz, and T. Vicsek. Deterministic scale-free networks. *Physica A*, 299(3-4):559–564, 2001.

[132] C. Song, L. K. Gallos, S. Havlin, and H. A. Makse. How to calculate the fractal dimension of a complex network: The box covering algorithm. *J. Stat. Mech.*, 2007(3):P03006, 2007.

[133] G. Bizhani, V. Sood, M. Paczuski, and P. Grassberger. Random sequential renormalization of networks: Application to critical trees. *Phys. Rev. E*, 83(3):036110, 2011. doi: 10.1103/PhysRevE.83.036110.

[134] S.-W. Son, G. Bizhani, C. Christensen, P. Grassberger, and M. Paczuski. Irreversible aggregation and network renormalization. *Europhys. Lett.*, 95(5):58007, 2011.

[135] Z. Zhang, S. Zhou, T. Zou, and G. Chen. Fractal scale-free networks resistant to disease spread. *J. Stat. Mech.*, 2008(9):P09008, 2008.

[136] G. Csányi and B. Szendrői. The fractal–small-world dichotomy in real-world networks. *Phys. Rev. E*, 70(1-2):016122, 2004.

[137] F. Kawasaki and K. Yakubo. Reciprocal relation between the fractal and the small-world properties of complex networks. *Phys. Rev. E*, 82(3):036113, 2010.

[138] F. Radicchi, A. Barrat, S. Fortunato, and J. J. Ramasco. Renormalization flows in complex networks. *Phys. Rev. E*, 79(2):026104, 2009.

[139] M. Á. Serrano, D. Krioukov, and M. Boguñá. Percolation in self-similar networks. *Phys. Rev. Lett.*, 106(4):048701, 2011.

[140] A. Zeng and L. Lu. Coarse graining for synchronization in directed networks. *Phys. Rev. E*, 83(5):056123, 2011.

[141] N. Christofides. An algorithm for the chromatic number of a graph. *Computer J.*, 14: 38–39, 1971.

[142] S. Wilf. Backtrack: An O(1) expected time algorithm for the graph coloring problem. *Inf. Proc. Lett.*, 18:119–121, 1984.

[143] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2001.

[144] J. S. Kim, K. I. Goh, B. Kahng, and D. Kim. A box-covering algorithm for fractal scaling in scale-free networks. *Chaos*, 17(2):026116, 2007.

[145] L. C. Freeman. Centered graphs and the structure of ego networks. *Math. Soc. Sci.*, 3 (3):291–304, 1982.

[146] M. Everett and S. P. Borgatti. Ego network betweenness. *Soc. Networks*, 27(1):31–38, 2005.

[147] G. Tibély, L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki. Communities and beyond: Mesoscopic analysis of a large social network with complementary methods. *Phys. Rev. E*, 83(5):056125, 2011.

[148] L. Hébert-Dufresne, A. Allard, V. Marceau, P. A. Noël, and L. J. Dubé. Structural preferential attachment: Network organization beyond the link. *e-print arXiv:1105598ov2*, 2011.

[149] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76(3):036106, 2007.

[150] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[151] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, 2004.

[152] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *P. Natl. Acad. Sci. USA*, 104(1):36–41, 2007.

[153] B. H. Good, Y. A. De Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81(4): 046106, 2010.

[154] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74(3):036104, 2006.

[155] B. Bollobás. *Modern graph theory*. Springer, 1998.

[156] P. Erdős and A. Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[157] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.*, 54(4):396–405, 2003.

[158] V. Batagelj and Mrvar A. Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/, 2011.

[159] L. Šubelj. Network data. http://lovro.lpt.fri.uni-lj.si/?navigation=research$_$support, 2011.

[160] B. Jones. Computational geometry. http://compgeom.cs.uiuc.edu/~jeffe/compgeom/, 2002.

[161] L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the KDD Workshop on Link discovery*, pages 36–43. ACM, 2005.

[162] VAST Challenge '08. http://www.cs.umd.edu/hcil/VASTchallenge08/, 2008.

[163] D. E. Knuth. *The Stanford GraphBase: A platform for combinatorial computing*. ACM Press, 1993.

[164] L. Šubelj and M. Bajec. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Phys. Rev. E*, 83(3):036103, 2011.

[165] J. M. Reitz. Online Dictionary of Library and Information Science. http://vax.wcsu.edu/library/odlis.html, 2002.

[166] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. University of South Florida free association norms. http://w3.usf.edu/FreeAssociation/, 2011.

[167] L. Šubelj and M. Bajec. Community structure of complex software systems: Analysis and applications. *Physica A*, 390(16):2968–2975, 2011.

[168] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.*, 31(9):2443–2450, 2003.

[169] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80(5):056117, 2009.

[170] L. C. Freeman. Centrality in social networks conceptual clarification. *Soc. networks*, 1(3): 215–239, 1979.

[171] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A*, 390(6): 1150–1170, 2011.

[172] L. S. Heath and N. Parikh. Generating random graphs with tunable clustering coefficient. *Physica A*, 390(23-24):4577–4587, 2011.

[173] H. Park and S. Moon. Sampling bias in user attribute estimation of osns. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 183–184. International World Wide Web Conferences Steering Committee, 2013.

[174] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie. Sampling biases in ip topology measurements. In *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications*, volume 1, pages 332–341. IEEE, 2003.

[175] A. S. Maiya and T. Y. Berger-Wolf. Benefits of bias: towards better characterization of network sampling. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 105–113. ACM, 2011.

[176] B. Long, X. Wu, Z. Zhang, and P. S. Yu. Community learning by graph approximation. In *Proceedings of 7th IEEE International Conference on Data Mining*, pages 232–241. IEEE, 2007.

[177] F. Wu and B. A. Huberman. Finding communities in linear time: a physics approach. *Eur. Phys. J. B*, 38(2):331–338, 2004.

[178] M. Rosvall and C. T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *P. Natl. Acad. Sci. USA*, 104(18): 7327–7331, 2007.

[179] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *P. Natl. Acad. Sci. USA*, 101(9):2658–2663, 2004.

[180] J. Reichardt and D. R. White. Role models for complex networks. *Eur. Phys. J. B*, 60(2): 217–224, 2007.

[181] B. Lužar, Z. Levnajić, J. Povh, and M. Perc. Community structure and the evolution of interdisciplinarity in slovenia's scientific collaboration network. *PloS One*, 9(4):e94429, 2014.

[182] M. Perc. The matthew effect in empirical data. *J. Roy. Soc. Interface*, 11(98):20140378, 2014.

[183] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks ISDN Syst.*, 30(1): 107–117, 1998.

[184] N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: from static to streaming graphs. *e-print arXiv:11211.3412*, 2012.

[185] L. A. Goodman. Snowball sampling. *Ann. Math. Stat.*, pages 148–170, 1961.

[186] Y. Zhao, E. Levina, and J. Zhu. Community extraction for social networks. *P. Natl. Acad. Sci.*, 108(18):7321–7326, 2011.

[187] L. Šubelj, S Žitnik, N. Blagus, and M. Bajec. Node mixing and group structure of complex software neworks. *Advs. Complex Syst.*, 17: 1450022, 2014.

[188] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bull. Soc. Vaud. Sci. Nat.*, 37:547–579, 1901.

[189] S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach (second edition)*. Upper Saddle River, N. J.: Prentice Hall, 2003.

[190] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[191] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[192] M. E. J Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, 2003.

[193] L. Šubelj, D. Fiala, and M. Bajec. Network-based statistical comparison of citation topology of bibliographic databases. *Sci. Rep.*, 4:6496, 2014.

[194] R. D. Cook and S. Weisberg. Residuals and influence in regression. *Mg. Stat. Pro.*, 1982.

[195] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th international conference on World Wide Web*, pages 114–118. ACM, 2001.

[196] N. K. Ahmed, J. Neville, and R. Kompella. Reconsidering the foundations of network sampling. In *Proceedings of the 2nd Workshop on Information in Networks*, 2010.

[197] J. Reimand, L. Tooming, H. Peterson, P. Adler, and J. Vilo. Graphweb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic acids research*, 36(suppl 2):W452–W459, 2008.

[198] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.